SPECIAL COLLECTOR'S EDITION

SCIENTIFIC AMERICAN

PLUS 4 UNSOLVED SCIENTIFIC MYSTERIES

REVOLUTIONARY SCIENCE

Discoveries that are changing our understanding of the world, from biology to physics

INCLUDING

- How the brain makes reality
- Dinosaurs' true colors
- Origin of the universe in question
- The truth about human metabolism
- Viking women's crucial role in ancient economy
- Violent collisions created our solar system

WINTER 2023

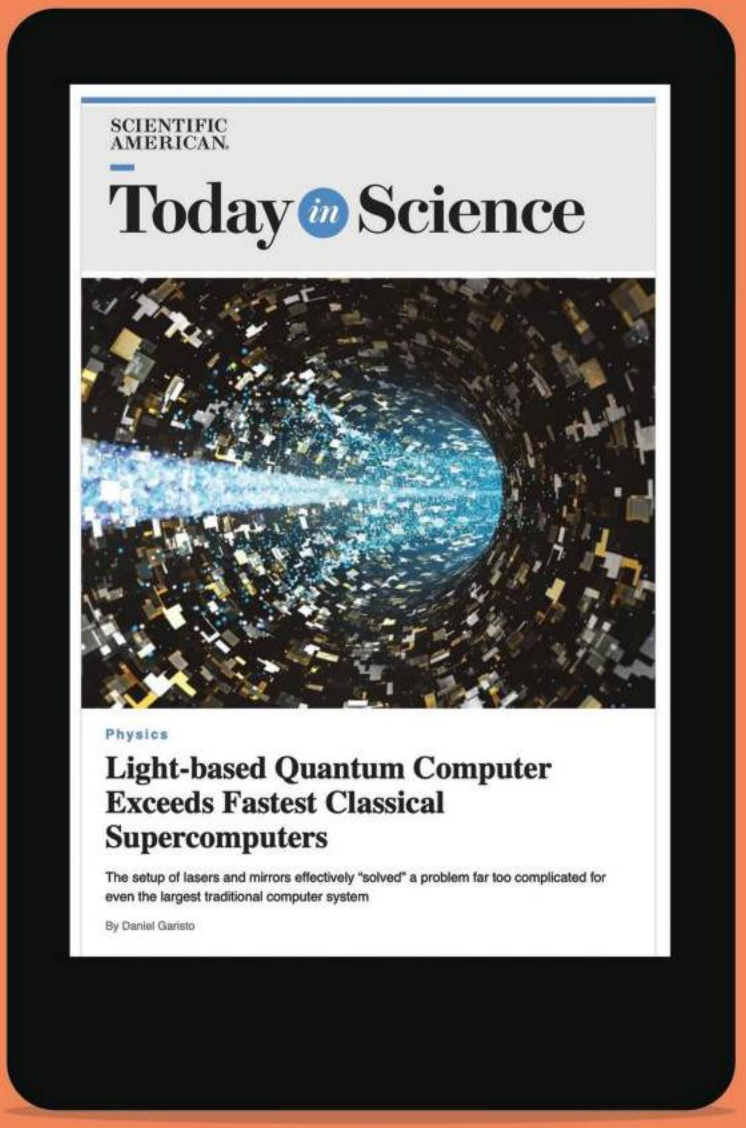
© 2023 Scientific American

SCIENTIFIC AMERICAN

Science in your inbox

Sign up for our free daily newsletter

Scan to learn more


SCIENTIFIC AMERICAN
Today in Science
Physics
Light-based Quantum Computer Exceeds Fastest Classical Supercomputers
The setup of lasers and mirrors effectively "solved" a problem far too complicated for even the largest traditional computer system
By Daniel Garisto

Scientific American is a registered trademark of Springer Nature America, Inc.

SCIENTIFIC AMERICAN
ESTABLISHED 1845

Revolutionary Science is published by the staff of Scientific American, with project management by:

Editor in Chief: **Laura Helmuth**

Managing Editor: **Jeanna Bryner**

Chief Newsletter Editor: **Andrea Gawrylewski**

Creative Director: **Michael Mrak**

Issue Designer: **Lawrence R. Gendron**

Senior Graphics Editor: **Jen Christiansen**

Associate Graphics Editor: **Amanda Montañez**

Photography Editor: **Monica Bradley**

Issue Photo Researcher: **Beatrix Mahd Soltani**

Copy Director: **Maria-Christina Keller**

Senior Copy Editors: **Angelique Rondeau, Aaron Shattuck**

Managing Production Editor: **Richard Hunt**

Prepress and Quality Manager: **Silvia De Santis**

Executive Assistant Supervisor: **Maya Harty**

Senior Editorial Coordinator: **Brianne Kane**

President: **Kimberly Lau**

Publisher and Vice President: **Jeremy A. Abbate**

Vice President, Commercial: **Andrew Douglas**

Vice President, Content Services: **Stephen Pincock**

Associate Vice President, Business Development: **Diane McGarvey**

Programmatic Product Manager: **Zoya Lysak**

Product Managers: **Ian Kelly, Miguel Olivares**

Head, Marketing: **Christopher Monello**

Senior Commercial Operations Coordinator: **Christine Kaelin**

Custom Publishing Editor: **Lisa Pallatroni**

Press Manager: **Sarah Hausman**

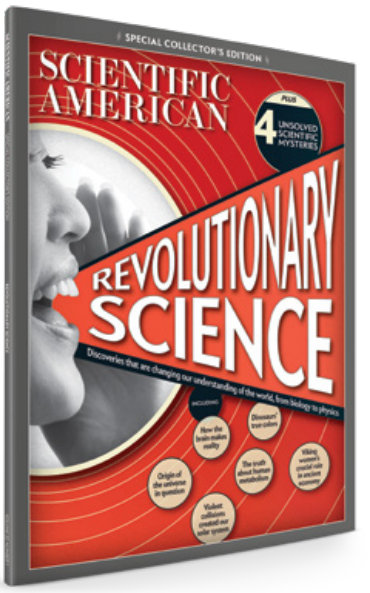
Production Controller: **Madelyn Keyes-Milch**

Advertising Production Manager: **Michael Broomes**

JGI/Getty Images (*portrait*); *Scientific American* (*cover design*)

FROM THE EDITOR

Science Out of the Box



On balance, science tends to follow a slow and incremental path: new observations build on previous findings, often over many years, contributing piecemeal to a larger body of knowledge. But every now and then, insights arise that can dramatically accelerate our understanding of a given subject or take the field in unexpected directions. In this special edition, we've rounded up just these kinds of advances—what Thomas S. Kuhn called "paradigm shifts" in his 1962 book *The Structure of Scientific Revolutions*.

You'll see in this anthology that revolutionary discoveries can happen in nearly every field. In biology and evolution, researchers have uncovered a remarkable sense of magnetism within birds' eyes that they use to navigate during migrations (*page 4*). Dinosaurs, long thought of and depicted as impressive but drab-looking creatures, in fact weren't drab at all—research on fossilized melanin cells shows that dinosaurs were feathered in bright colors; some even sported dots and stripes (*page 10*). In human history, archaeologists are amending their assumptions about the economic roles that women played in Viking and medieval society, based on recent analyses of North Atlantic textiles (*page 26*).

In cosmology, data from NASA's transformative new observatory—the James Webb Space Telescope—threaten to upend astrophysicists' narrative about how the universe formed (*page 42*), and physicists have devised some astounding ideas about entangled black holes that will help explain an almost century-old paradox (*page 52*). Current views suggest that even our solar system did not establish itself slowly, as was long thought, but was born of a violent blur of impacts and destruction (*page 34*). All of which is confounded by the possibility that what we "see" out in space may not actually be there, in the quantum sense (*page 60*).

Major advances in our understanding of our own minds and bodies are just as striking. Trauma from events experienced during pregnancy can be passed down to children, evidenced in genetic changes (*page 72*). Evolutionary anthropologist Herman Pontzer and his colleagues have painstakingly gathered surprising information about metabolism and how our bodies burn calories (*page 66*). And the mRNA vaccine technologies developed during the COVID pandemic are poised to revolutionize treatment for a slew of diseases, from herpes simplex virus 2 to norovirus and rabies, to name a few (*page 78*).

One of my graduate school advisers used to say that what humans Know (capital "K") is infinitesimal compared with what is knowable in the universe. In that spirit, we also present a sample of some of the most fascinating unanswered questions, from the enduring mystery of how human consciousness arises (*page 104*) to a growing debate over the very nature of the cosmos (*page 88*).

Some of the most astonishing findings collected here are those where conventional scientific practice went out the window. As lichenologist Toby Spribille tells reporter Erica Gies in the story on page 18, "You have this culture of prepared minds that makes it extremely difficult to think outside the box. It creates the box." Indeed, the beauty of scientific inquiry is that when we quickly leapfrog the pace of discovery—or take our explorations in totally novel directions—we uncover an entirely new way of thinking. Here's to breaking the box. SA

Andrea Gawrylewski
Chief Newsletter Editor,
editors@sciam.com

© 2023 Scientific American

SPECIAL EDITION

SCIENTIFIC AMERICAN

Volume 32, Number 1, Winter 2023

BIOLOGY & EVOLUTION

4 The Quantum Nature of Bird Migration
Migratory birds travel vast distances between their breeding and wintering grounds. Recent research hints at the biophysical underpinnings of their internal navigation system. *By Peter J. Hore and Henrik Mouritsen*

10 The True Colors of Dinosaurs
Long thought impossible, preservation of fossil pigments is allowing scientists to reconstruct extinct organisms with unprecedented accuracy—a feat that is yielding surprising insights into the lives they led. *By Jakob Vinther*

18 The Meaning of Lichen
How a naturalist's observations in the wilds of British Columbia inspired a scientist to discover hidden symbioses—overturning 150 years of accepted scientific wisdom. *By Erica Gies*

26 The Power of Viking Women
Analyses of ancient North Atlantic textiles show that Viking and medieval women wielded considerable cultural and economic influence. *By Francine Russo*

SPACE & PHYSICS

34 Solar System Smashup
Our neighborhood of planets was not created slowly, as scientists once thought, but in a speedy blur of high-energy crashes, destruction and rebuilding. *By Linda T. Elkins-Tanton*

42 Breaking Cosmology
JWST's first images include unimaginably distant galaxies that challenge theories of how quickly these structures can form. *By Jonathan O'Callaghan*

52 Black Holes, Wormholes and Entanglement
Researchers cracked a paradox by considering what happens when the insides of black holes are connected by spacetime wormholes. *By Ahmed Almheiri*

60 The Universe Is Not Locally Real
Experiments with entangled light have revealed a profound mystery at the heart of reality. *By Daniel Garisto*

BODY & MIND

66 The Human Engine
Studies of metabolism reveal surprising insights into how we burn calories—and how cooperative food production helped *Homo sapiens* flourish. *By Herman Pontzer*

72 Trauma in the Family Tree
Parents' adverse experiences leave biological traces in children. *By Rachel Yehuda*

78 Messenger RNA Therapies Finally Arrive
Instructing our cells to make specific proteins could control influenza, autoimmune diseases, even cancer. *By Drew Weissman*

80 Constructing the World from Inside Out
The brain probes your physical surroundings to select just the information needed to survive and flourish. *By György Buzsáki*

UNSOLVED QUESTIONS

88 Twin Tensions
A debate over conflicting measurements of key cosmological properties is poised to shape the next decade of astronomy and astrophysics. *By Anil Ananthaswamy*

94 Wonder of the Ancient World
Scientists have revealed intriguing details about the complex gearing of the Antikythera mechanism of ancient Greece. *By Tony Freeth*

104 Our Inner Universes
Reality is constructed by the brain, and no two brains are exactly alike. *By Anil K. Seth*

110 Long Haulers and Chronic Illnesses
Millions of Americans have long COVID—and we are not prepared to manage the crisis. *By Meghan O'Rourke*

DEPARTMENTS

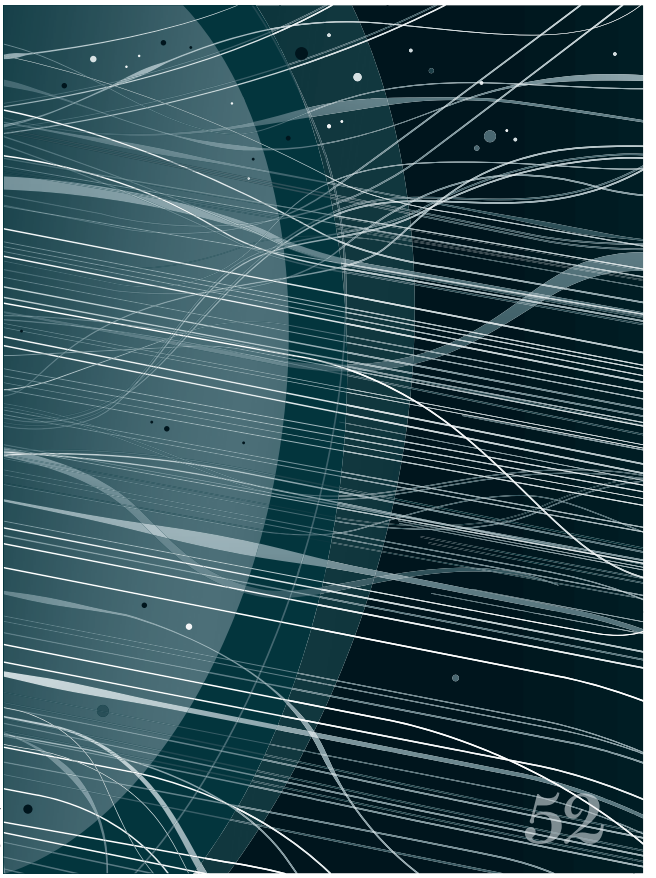
1 FROM THE EDITOR
Science Out of the Box

112 END NOTE
Sleeping Beauties of Science
Some of the best research can slumber for years. *By Amber Williams*

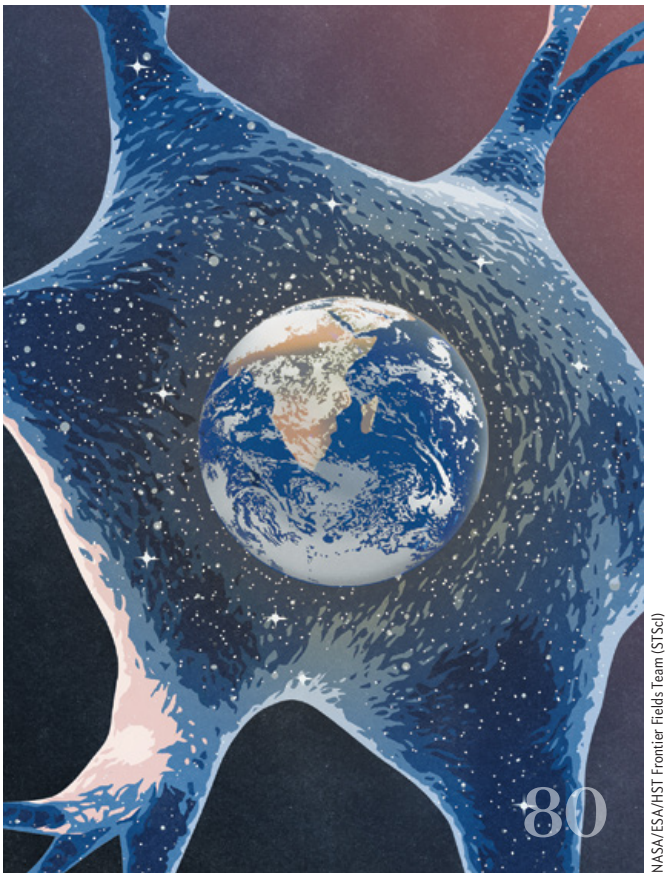
Articles in this special issue are updated or adapted from previous issues of *Scientific American* and from ScientificAmerican.com. Copyright © 2023 Scientific American, a division of Springer Nature America, Inc. All rights reserved. Scientific American Special (ISSN 1936-1513), Volume 32, Number 1, Winter 2023, published by Scientific American, a division of Springer Nature America, Inc., 1 New York Plaza, Suite 4600, New York, N.Y. 10004-1562. Canadian BN No. 127387652RT; TVQ1218059275 TQ0001. To purchase additional quantities: U.S., $13.95 each; elsewhere, $17.95 each. Send payment to Scientific American Back Issues, P.O. Box 3187, Harlan, Iowa 51537. **Inquiries: fax 212-355-0408 or telephone 212-451-8415. Printed in U.S.A.**

© 2023 Scientific American

Chase Stone



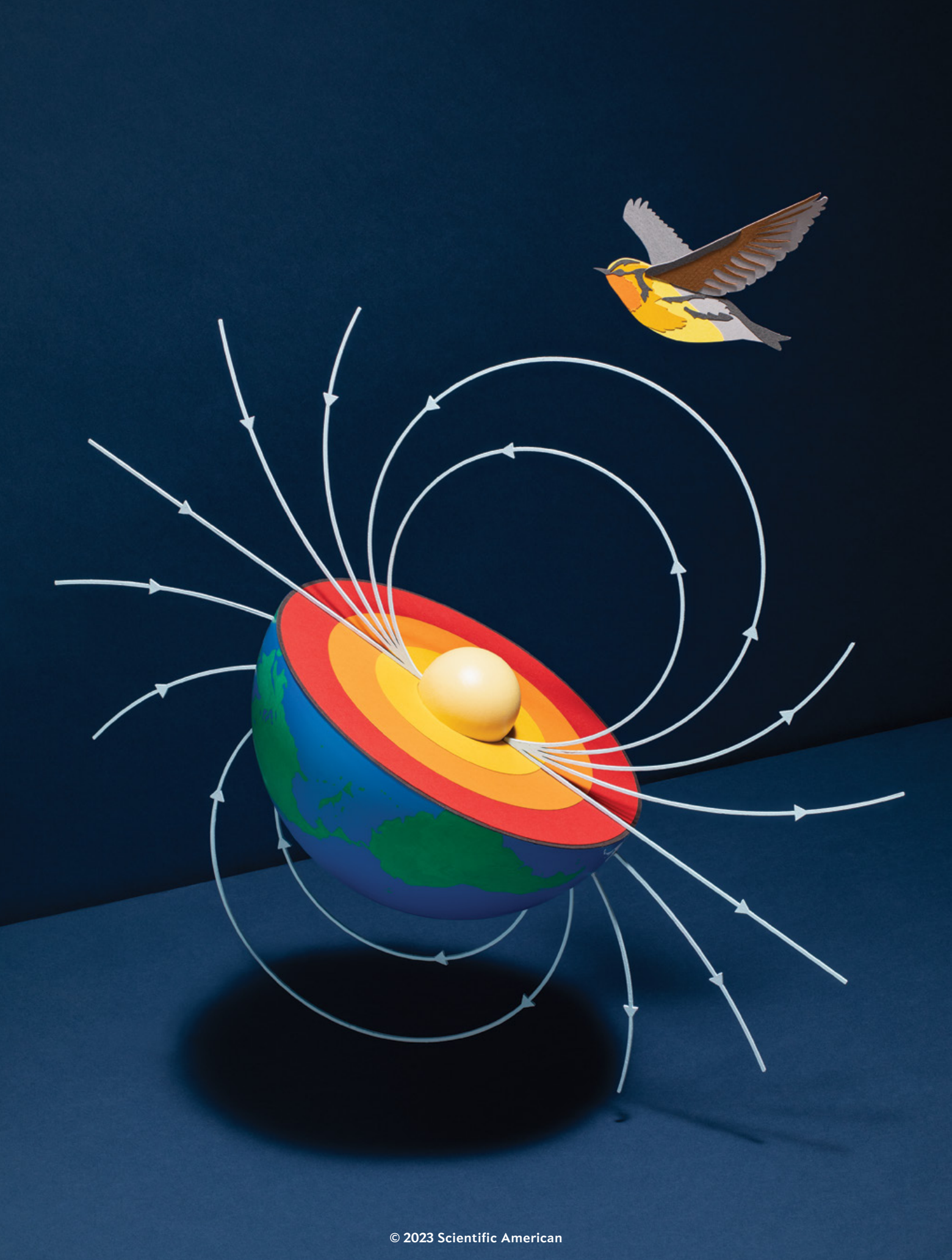
Harry Campbell



Stefania Infante

NASA/ESA/HST Frontier Fields Team (STScI)

© 2023 Scientific American



© 2023 Scientific American

The Quantum Nature of Bird Migration

Migratory birds travel vast distances between their breeding and wintering grounds. Recent research hints at the biophysical underpinnings of their internal navigation system

By Peter J. Hore and Henrik Mouritsen

Illustration by Kyle Bean

© 2023 Scientific American

IMAGINE YOU ARE A YOUNG BAR-TAILED GODWIT, A LARGE, LEGGY SHOREBIRD WITH a long, probing bill hatched on the tundra of Alaska. As the days become shorter and the icy winter looms, you feel the urge to embark on one of the most impressive migrations on Earth: a nonstop transequatorial flight lasting at least seven days and nights across the Pacific Ocean to New Zealand 12,000 kilometers away. It's do or die. Every year tens of thousands of Bar-tailed Godwits complete this journey successfully. Billions of other young birds, including warblers and flycatchers, terns and sandpipers, set out on similarly spectacular and dangerous migrations every spring, skillfully navigating the night skies without any help from more experienced birds.

People have long puzzled over the seasonal appearances and disappearances of birds. Aristotle thought that some birds such as swallows hibernated in the colder months and that others transformed into different species—redstarts turned into robins for the winter, he proposed. Only in the past century or so, with the advent of bird banding, satellite tracking and more widespread field studies, have researchers been able to connect bird populations that winter in one area and nest in another and show that some travel vast distances between the two locales every year. Remarkably, even juvenile long-haul travelers know where to go, and birds often take the same routes year after year. How do they find their way?

Migrating birds use celestial cues to navigate, much as sailors of yore used the sun and stars to guide them. But unlike humans, birds also detect the magnetic field generated by Earth's molten core and use it to determine their position and direction. Despite more than 50 years of research into magnetoreception in birds, scientists have been unable to work out exactly how they use this information to stay on course. Recently we and others have made inroads into this enduring mystery. Our experimental evidence suggests something extraordinary: a bird's compass relies on subtle, fundamentally quantum effects in short-lived molecular fragments, known as radical pairs, formed photochemically in its eyes. That is, the creatures appear to be able to "see" Earth's magnetic field lines and use that information to chart a course between their breeding and wintering grounds.

A MYSTERIOUS SENSE

MIGRATORY BIRDS have an internal clock with an annual rhythm that tells them, among other things, when to migrate. They also inherit from their parents the directions in which they need to fly in the autumn and spring, and if the parents each have different genetically encoded directions, their offspring will end up with an intermediate direction. For example, if a southwest-migrating bird is crossed with a southeast-migrating bird, their offspring will head south when the time comes. But how do the young birds know which direction is southwest or south or southeast? They have at least three different compasses at their disposal: one allows them to extract information from the position of the sun in the sky, another uses the patterns of the stars at night, and the third is based on Earth's ever present magnetic field.

In their first autumn, young birds follow inherited instructions such as "fly southwest for three weeks and then south-southeast for two weeks." If they make a mistake or are blown off course, they are generally unable to recover because they do not yet have a functional map that would tell them where they are. This is one of the reasons that only 30 percent of small songbirds survive their first migrations to their wintering grounds and back again. During its first migration a bird builds a map in its brain that, on subsequent journeys, will enable it to navigate with an ultimate precision of centimeters over thousands of kilometers. Some birds breed in the same nest box and sleep on the same perch in their wintering range year after year. Equipped with this map, about 50 percent of adult songbirds make it back to their nesting site to breed every year.

Migratory birds' navigational input comes from several senses—mainly sight, smell and magnetoreception. By observing the apparent nighttime rotation of the stars around the North Star, the birds learn to locate north before they embark on their first migration, and an internal 24-hour clock allows them to calibrate their sun compass. Characteristic smells can help birds recognize places they have visited before. Scientists know a great deal about the detailed biophysical mechanisms of the birds' senses of sight and smell. But the inner workings of their magnetic compass have proved harder to understand.

The magnetic direction sense in small songbirds that migrate at night is remarkable in several important respects. First, observations of caged birds exposed to carefully controlled magnetic fields show that their compass does not behave like the magnetized needle in a ship's compass. A bird detects the axis of the magnetic field and the angle it makes with Earth's surface, the so-called inclination compass. In laboratory experiments, inverting the magnetic field's direction so it points in exactly the opposite direction has no effect on the bird's ability to orient correctly. Second, a bird's perception of Earth's magnetic field can be disrupted by extraordinarily weak magnetic fields that reverse their direction several million times per second. Last, even though songbirds fly at night under the dim light of the stars, their magnetic compass is light-dependent, hinting at a link between vision and magnetic sensing.

In 1978, in an attempt to make sense of these features of avian magnetoreception, the late Klaus Schulten, then at the Max Planck Institute for Biophysical Chemistry in Göttingen, Germany, put forth a remarkable idea: the compass relies on magnetically sensitive chemical transformations. At first glance, this proposal seems preposterous because the energy available from Earth's magnetic field is millions of times too small to break, or even significantly weaken, the bonds between atoms in molecules. But Schulten was inspired by the discovery 10 years previously that short-lived chemical intermediates known as radical pairs have unique properties that make their chemistry sensitive to feeble magnetic interactions. Over the past 40 years researchers have conducted hundreds of lab studies of radical-pair reactions that are affected by magnetic fields.

To appreciate why radical pairs are so special, we need to talk about a quantum-mechanical property of the electron known as

© 2023 Scientific American

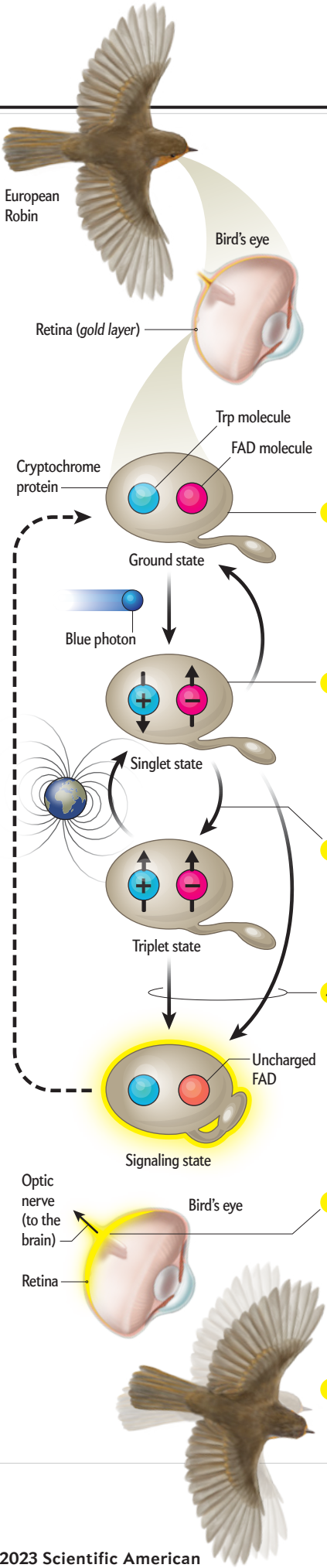
spin angular momentum, or "spin" for short. Spin is a vector with a direction as well as a magnitude, and it is often represented by an arrow, ↑ or ↓, for example. Particles with spin have magnetic moments, which is to say they behave like microscopic magnets. Most molecules have an even number of electrons arranged in pairs with opposed spins (↑↓), which therefore cancel each other out. Radicals are molecules that have lost or gained an electron, meaning they contain an odd, unpaired, electron and hence have a spin and a magnetic moment. When two radicals are created simultaneously by a chemical reaction (this is what we mean by radical pair), the two unpaired electrons, one in each radical, can have either antiparallel spins (↑↓) or parallel spins (↑↑), arrangements known as singlet and triplet states, respectively.

Immediately after a radical pair is created in a singlet state, internal magnetic fields cause the two electronic spins to undergo a complex quantum "waltz" in which singlet turns into triplet and triplet turns back into singlet millions of times per second for periods of up to a few microseconds. Crucially, under the right conditions, this dance can be influenced by external magnetic fields. Schulten suggested that this subtle quantum effect could form the basis of a magnetic compass sense that might respond to environmental stimuli a million times weaker than would normally be thought possible. Research that we and others have carried out in recent years has generated fresh support for this hypothesis.

To be useful, hypotheses need to explain known facts and make testable predictions. Two aspects of Schulten's proposed compass mechanism are consistent with what is known about the birds' compass: radical pairs are indifferent to exact external magnetic field reversals, and radical pairs are often formed when molecules absorb light. Given that the birds' magnetic compass is light-dependent, a prediction of Schulten's hypothesis is that their eyes play a part in the magnetic sensory system. About 10 years ago the research group of one of us (Mouritsen) at the University of Oldenburg in Germany found that a brain region called Cluster N, which receives and processes visual information, is by far the most active part of the brain when certain night-migrating birds are using their magnetic compass. If Cluster N is dysfunctional, research in migratory European Robins showed, the birds can still use their sun and star compasses,

The Compass Mechanism

Studies suggest that the magnetic compass of migratory birds relies on quantum effects in short-lived molecular fragments known as radical pairs that are formed photochemically in the eyes. In this way, the birds can perceive Earth's magnetic field lines and use that information to navigate their long-haul trips.



1. Cryptochrome proteins—located in the retinal cells of the bird's eye—include a flavin adenine dinucleotide molecule (FAD) and a tryptophan amino acid (Trp). In the stable state, these molecules are electrically neutral, and a small section of the protein probably extends like a tail.

2. When a photon of blue light hits the cryptochrome, an electron jumps from the Trp onto the FAD. The resulting molecules—each with an odd number of electrons—are known as a radical pair. In this singlet state, the molecules' unpaired electrons spin in opposition.

3. The activated protein oscillates rapidly back and forth between the singlet state and the triplet state, in which the unpaired electrons spin in parallel. Earth's magnetic field influences the spin, impacting the likelihood of each state dominating.

4. Both states can undergo chemical reactions that transform them into the "signaling state"—in which a hydrogen ion has been added to the FAD radical—and the tail seems to move closer to the body of the protein. The singlet state can also simply return to the ground state. The proportion of outcomes depends on the bird's orientation in Earth's magnetic field.

5. The signaling state of the cryptochrome turns on a biochemical cascade that triggers the release of neurotransmitter molecules in the retina. Signals continue to the bird's brain, where the magnetic information they contain is integrated with information from other directional cues, informing the direction of the bird's flight.

6. Cryptochrome returns to its ground state, and the process starts again (*dashed arrow*).

Graphic by Jillian Ditner (birds and eyes) and Jen Christiansen

© 2023 Scientific American

but they are incapable of orienting using Earth's magnetic field. From experiments such as these, it is clear that the magnetic compass sensors are located in the birds' retinas.

One early objection to the radical-pair hypothesis was that no one had ever shown that magnetic fields as tiny as Earth's, which are 10 to 100 times weaker than a fridge magnet, could affect a chemical reaction. To address this point, Christiane Timmel of the University of Oxford and her colleagues chose a molecule chemically unlike anything one would find inside a bird: one that contained an electron using Earth's magnetic field. From experiments such as these, it is clear that the magnetic compass sensors are located in the birds' retinas. One early objection to the radical-pair hypothesis was that no one had ever shown that magnetic fields as tiny as Earth's, which are 10 to 100 times weaker than a fridge magnet, could affect a chemical reaction. To address this point, Christiane Timmel of the University of Oxford and her colleagues chose a molecule chemically unlike anything one would find inside a bird: one that contained an electron-donor molecule linked to an electron-acceptor molecule via a molecular bridge. Exposing the molecules to green light caused an electron to jump from the donor to the acceptor over a distance of about four nanometers. The radical pair that formed from this reaction was extremely sensitive to weak magnetic interactions, proving that it is indeed possible for a radical-pair reaction to be influenced by the presence of—and, more important, the direction of—an Earth-strength magnetic field.

EUROPEAN ROBIN (*top*) and Bar-tailed Godwit (*bottom*) are among the many birds that migrate long distances.

Schulten's hypothesis also predicts that there must be sensory molecules (magnetoreceptors) in the retina in which magnetically sensitive radical pairs can be created using the wavelengths birds need for their compass to operate, which another line of research had identified as light centered in the blue region of the spectrum. In 2000 he suggested that the necessary photochemistry could take place in a then recently discovered protein called cryptochrome.

Cryptochromes are found in plants, insects, fish, birds and humans. They have a variety of functions, including light-dependent control of plant growth and regulation of circadian clocks. What makes them attractive as potential compass sensors is that they are the only known naturally occurring photoreceptors in any vertebrate that form radical pairs when they absorb blue light. Six types of cryptochromes have been found in the eyes of migratory birds, and no other type of candidate magnetoreceptor molecule has emerged in the past 20 years.

Like other proteins, cryptochromes are composed of chains of amino acids folded up into complex three-dimensional structures. Buried deep in the center of many cryptochromes is a yellow molecule called flavin adenine dinucleotide (FAD) that, unlike the rest of the protein, absorbs blue light. Embedded among the 500 or so amino acids that make up a typical cryptochrome is a roughly linear chain of three or four tryptophan amino acids stretching from the FAD out to the surface of the protein. Immediately after the FAD absorbs a blue photon, an electron from the nearest tryptophan hops onto the flavin portion of the FAD. The first tryptophan then attracts an electron from the second tryptophan and so on. In this way, the tryptophan chain behaves like a molecular wire. The net result is a radical pair made of a negatively charged FAD radical in the center of the protein and, two nanometers away, a positively charged tryptophan radical at the surface of the protein.

In 2012 one of us (Hore), working with colleagues at Oxford, carried out experiments to test the suitability of cryptochrome as a magnetic sensor. The study used cryptochrome-1, a protein found in *Arabidopsis thaliana,* the plant in which cryptochromes had been discovered 20 years earlier. Using short laser pulses to produce radical pairs inside the purified proteins, we found that we could fine-tune their subsequent reactions by applying magnetic fields. This was all very encouraging, but, of course, plants don't migrate.

We had to wait almost a decade before we could make similar measurements on a cryptochrome from a migratory bird. The first challenge was to decide which of the six bird cryptochromes to look at. We chose cryptochrome-4a (Cry4a), partly because it binds FAD much more strongly than do some of its siblings, and if there is no FAD in the protein, there will be no radical pairs and no magnetic sensitivity. Experiments in Oldenburg also showed that the levels of Cry4a in migratory birds are higher during the spring and autumn migratory seasons than they are during winter and summer when the birds do not migrate. Computer simulations performed by Ilia Solov'yov in Oldenburg showed that European Robin Cry4a has a chain of four tryptophans—one more than the Cry1 from *Arabidopsis.* Naturally, we wondered whether the extended chain had evolved to optimize magnetic sensing in migratory birds.

Our next challenge was to get large amounts of highly pure robin Cry4a. Jingjing Xu, a Ph.D. student in Mouritsen's lab, solved it. After optimizing the experimental conditions, she was able to use bacterial cell cultures to produce samples of the protein with the FAD correctly bound. She also prepared versions of the protein in which each of the four tryptophans was replaced, one at a time, by a different amino acid so as to block electron hopping at each of the four positions along the chain. Working with these alternative

Adrian Coleman/Getty Images (*top*); Imogen Warren/Getty Images (*bottom*)

© 2023 Scientific American

versions of the protein would allow us to test whether the electrons are really jumping all the way along the tryptophan chain.

We shipped these samples—the first purified cryptochromes from any migratory animal—to Oxford, where Timmel and her husband, Stuart Mackenzie, studied them using the sensitive laser-based techniques they had developed specifically for that purpose. Their research groups found that both the third and fourth tryptophan radicals at the end of the chain are magnetically sensitive when paired with the FAD radical. We suspect that the tryptophans work cooperatively for efficient magnetic sensing, biochemical signaling and direction finding. We also speculate that the presence of the fourth tryptophan might enhance the initial steps of signal transduction, the process by which nerve impulses encoding the magnetic field direction are generated and ultimately sent along the optic nerve to the brain. We are currently conducting experiments to identify the proteins that interact with Cry4a.

One more cryptochrome finding deserves mention here. We compared robin Cry4a with the extremely similar Cry4a proteins from two nonmigratory birds, pigeons and chickens. The robin protein had the most magnetic sensitivity, hinting that evolution might have optimized robin Cry4a for navigation.

OPEN QUESTIONS

ALTHOUGH THESE EXPERIMENTS confirm that Cry4a has some of the properties required of a magnetoreceptor, we are still a long way from proving how migratory birds perceive Earth's magnetic field lines. One crucial next step is to determine whether radical pairs actually form in the eyes of migratory birds.

The most promising way to test for radical pairs inside the birds' eyes was inspired by the work of chemists and physicists who, in the 1980s, showed that fluctuating magnetic fields alter the way radical-pair reactions respond to static magnetic fields. Their work predicted that a weak radio-frequency electromagnetic field, fluctuating with the same frequencies as the "singlet-triplet waltz," might interfere with the birds' ability to use their magnetic compass. Thorsten Ritz of the University of California, Irvine, and his colleagues were the first to confirm this prediction in 2004.

In 2007 Mouritsen began similar behavioral experiments in his lab in Oldenburg—with intriguingly different results. During the spring and fall, birds that travel between nesting and wintering grounds exhibit a behavior called *Zugunruhe,* or migratory restlessness, as if they are anxious to get on their way. When caged, these birds usually use their magnetic compass to instinctively orient themselves in the direction in which they would fly in the wild. Mouritsen found that European Robins tested in wooden huts on his university's campus were unable to orient using their magnetic compass. He suspected that weak radio-frequency noise (sometimes called electrosmog) generated by electrical equipment in the nearby labs was interfering with the birds' magnetic compass.

To confirm that electrosmog was the source of the problem, Mouritsen and his team lined the huts with aluminum sheets to block the stray radio frequencies. On nights when the shields were grounded and functioned properly, the birds oriented well in Earth's magnetic field. On nights when the grounding was disconnected, the birds jumped in random directions. When tested in an unshielded wooden shelter typically used for horses some kilometers outside the city and well away from electrical equipment, the same birds had no trouble detecting the direction of the magnetic field.

These results are significant on several fronts. If the radio-frequency fields affect the magnetic sensor and not, say, some component of the signaling pathway that carries nerve impulses to the brain, then they provide compelling evidence that a radical-pair mechanism underpins the bird's magnetic compass. The main competing hypothesis, for which there is currently much less support, proposes that magnetic iron–containing minerals are the sensors. Any such particles that were large enough to align like a compass needle in Earth's magnetic field would be far too big to rotate in a much weaker field that reversed its direction millions of times per second. Furthermore, the radio-frequency fields that upset the birds' magnetic orientation are astonishingly weak, and we don't yet understand exactly how they could corrupt the directional information available from the much stronger magnetic field of Earth.

It is also remarkable that the birds in the Oldenburg lab were disoriented much more effectively by broadband radio-frequency noise (randomly fluctuating magnetic fields with a range of frequencies) than by the single-frequency fields mostly used by Ritz and his collaborators. We hope that by subjecting migratory songbirds to bands of radio-frequency noise with different frequencies we will be able to determine whether the sensors really are FAD-tryptophan radical pairs or whether, as some other investigators have suggested, another radical pair might be involved.

Many questions about the birds' magnetic compass remain, including whether the magnetic field effects on robin Cry4a observed in vitro also exist in vivo. We also want to see whether migratory birds with genetically suppressed Cry4a production are prevented from orienting using their magnetic compass. If we can prove that a radical-pair mechanism is behind the magnetic sense in vivo, then we will have shown that a biological sensory system can respond to stimuli several million times weaker than previously thought possible. This insight would enhance our understanding of biological sensing and provide new ideas for artificial sensors.

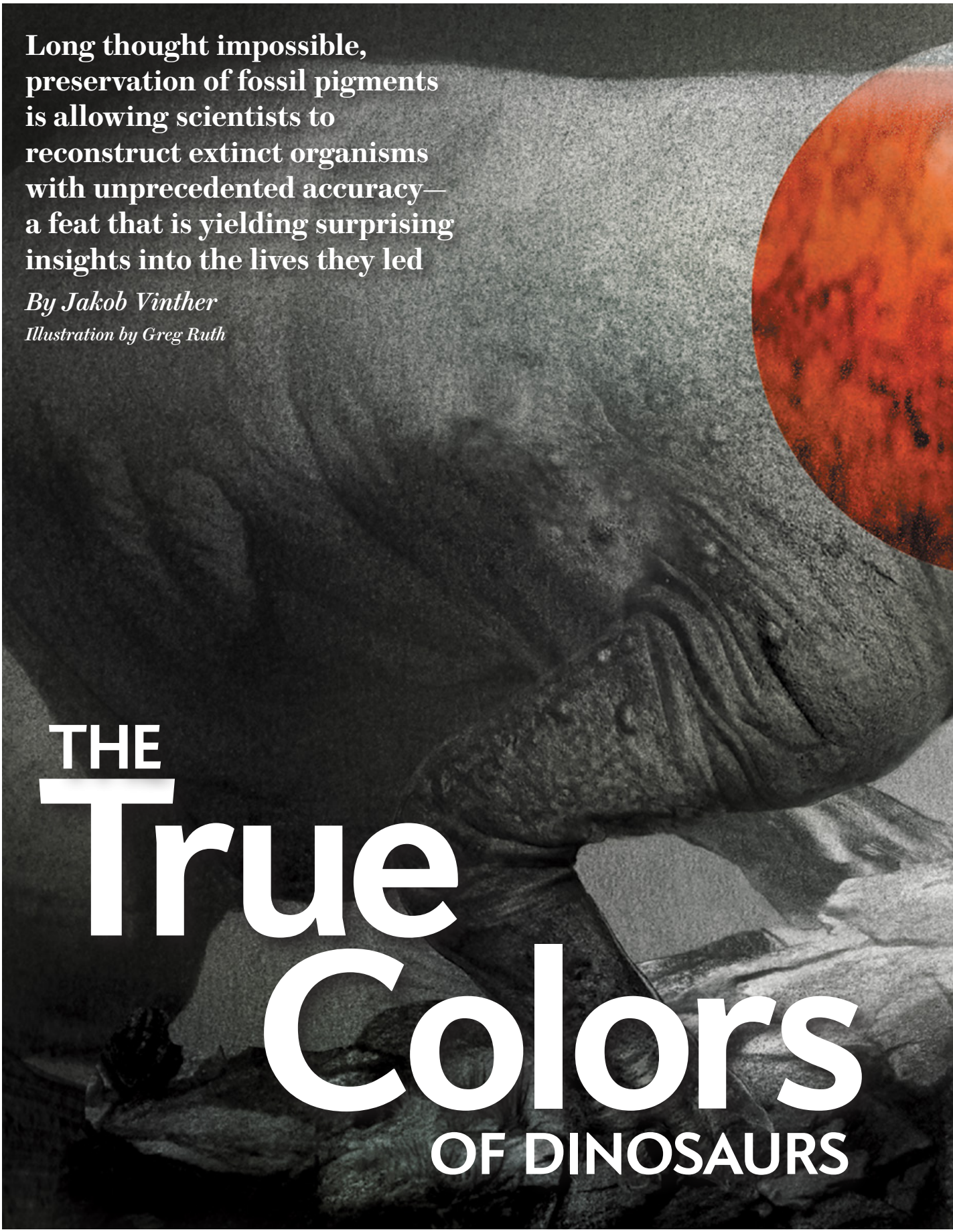
Working to gain a full understanding of the inner navigation systems of migratory birds is not merely an intellectual pursuit. One consequence of the enormous distances migratory birds travel is that they face more acute threats to their survival than most species that breed and overwinter in the same place. It is more difficult to protect them from the harmful effects of human activity, habitat destruction and climate change. Relocating migratory individuals away from damaged habitats is rarely successful because the birds tend to instinctively return to those unlivable locales. We hope that by providing new and more mechanistic insights into the ways in which these extraordinary navigators find their way, conservationists will have a better chance of "tricking" migrants into believing that a safer location really is their new home.

When you next see a small songbird, pause for a moment to consider that it might recently have flown thousands of kilometers, navigating with great skill using a brain weighing no more than a gram. The fact that quantum spin dynamics may have played a crucial part in its journey only compounds the awe and wonder with which we should regard these extraordinary creatures. SA

Peter J. Hore is a chemist at the University of Oxford. He works on the biophysical chemistry of electron and nuclear spins and their effects on processes such as the mechanisms of animal magnetoreception.

Henrik Mouritsen is a biologist at the University of Oldenburg in Germany. He studies the mechanisms of orientation and navigation in many different animals with an emphasis on night-migrating songbirds.

© 2023 Scientific American



Long thought impossible, preservation of fossil pigments is allowing scientists to reconstruct extinct organisms with unprecedented accuracy—a feat that is yielding surprising insights into the lives they led

By Jakob Vinther

Illustration by Greg Ruth

THE True Colors OF DINOSAURS

© 2023 Scientific American

PSITTACOSAURUS is one of several dinosaurs whose coloring has been deduced from fossilized pigments.

© 2023 Scientific American

ON A DAY IN OCTOBER 2006, I SAT IN A DARK LABORATORY at Yale University and zoomed in on the fossilized ink of a 200-million-year-old squid relative under an electron microscope. An ocean of translucent balls, each roughly a fifth of a micron in diameter, loomed into view. To the untrained eye, they might have been unimpressive. But I was riveted. These ancient structures looked exactly like the granules of melanin pigment that color the ink of modern squid and octopuses.

Perhaps I should not have been so surprised at the resemblance. Researchers had announced the first discovery of fossil ink granules a couple of years earlier. But seeing them with my own eyes was a revelation. As I examined cephalopod specimens from various locales and time periods, I realized their ink was always the same, perfectly preserved for hundreds of millions of years.

The consistently superb preservation of the ink made me wonder whether melanin might persist in fossils of other kinds of organisms. Melanin is the same pigment found in hair, skin, feathers and eyes. It can impart red, brown, gray and black hues and create metallic sheens. If I could find melanin in other fossils, perhaps I could reconstruct the coloring of extinct animals, including dinosaurs. For decades scientists assumed that pigments hardly ever survive the fossilization process. The few known examples all came from fossils of invertebrate creatures, not backboned ones. Thus, researchers could only guess at the colors of most long-vanished animals, using modern ones as a guide. As a result, dinosaur reconstructions have varied widely: some sport the drab earth tones associated with reptiles and amphibians; others flaunt the rainbow hues of modern birds (the only dinosaurs that have survived to modern times).

But discoveries I and others have made since the mid-2000s are taking out some of the guesswork. Our examinations of dozens of fossils have revealed many examples of melanin-bearing structures. By studying the shapes and organizations of these structures, we have been able to deduce the actual colors and patterns of extinct dinosaurs and other animals from deep time. These clues to the physical appearances of the creatures, in turn, have led to intriguing insights into their behaviors and habitats.

To test my hypothesis that melanin survives in other fossils and can be used to deduce the true colors of extinct animals, I wanted to find and analyze fossils with dark stains indicative of organic preservation in those anatomical regions generally known to contain melanin: the outer covering of the body and the eyes. And I needed to be able to examine the darkened areas under the electron microscope, which might require cutting a specimen down to size. Well-preserved fossils are rare, however, and museums guard them closely. Fortunately, a remarkable fossil site in my home country of Denmark called "Fur and Ølst Formation" had yielded exquisite bird fossils with feathers, which would be an ideal test case. I managed to convince the curator of vertebrate fossils at the Geological Museum in Copenhagen to cut down a typewriter-size block of limestone containing a skull of a little bird with stains where the eyes used to be and a dark halo of feather impressions into a piece the size of a slice of bread so that it could fit into the museum's electron microscope.

I had a good idea of what to search for under the microscope. Before obtaining the fossil bird for analysis, I had read numerous scientific papers to figure out what melanin looks like in the feathers of living birds. Melanin is synthesized in specialized cells known as melanocytes by cellular components called melanosomes. Typically the melanin remains encased in the melanosomes, which measure about 0.5 to two microns long and take two forms: a sausage-shaped kind that produces a form of melanin called eumelanin, which absorbs all wavelengths of light and thus gives squid ink and raven feathers their black color, and a meatball-shaped variant that makes pheomelanin, which imparts a rusty red hue. An absence of pigments results in white plumage. Gray and brown colors, for their parts, appear to arise from combinations of eumelanin, pheomelanin and pigment absence.

I had also consulted one of the world's leading experts on bird color: Richard Prum of Yale. Because I knew from the fossil

© 2023 Scientific American

ink that eumelanin could be preserved, I figured I would start by looking for that pigment in the feathers. Talking to Prum and his then Ph.D. student Vinod Saranathan, I learned that the sausage-shaped melanosomes line up in a distinctive way along the barbs and barbules that constitute a feather's branches. The melanosomes arrive there during development, when the melanocytes transfer them into specialized cells called keratinocytes that give rise to feathers and hair. If the dark stains on the feather impressions evident in the Danish bird fossil came from melanin, then I should see the sausages arranged this way along the feather branches under the microscope.

With great anticipation, I zoomed in on the fossil feathers—and encountered millions of sausage-shaped structures. Unfortunately, the local underground railway was less than 50 meters from the museum's basement, where the electron microscope was located; vibrations from the constant train traffic made it impossible to get a clear image. But the images were good enough to see the sausages. I immediately e-mailed them to my then Ph.D. supervisor at Yale, Derek Briggs, a pioneer in the study of extraordinarily preserved fossils. He replied with less enthusiasm than I had hoped for, noting that these structures were the same as those he and others had found in fossil feathers and mammal hair for decades and had identified as bacteria.

I still thought the sausages were melanosomes, though, and made my argument to Briggs. Not only did they have the right shape and size but their orientation in the feather structures mirrored that of black melanosomes in modern bird feathers. Furthermore, it was clear from the fossil squid ink that melanin can be fossilized. Briggs began to warm to the idea, but he was not convinced until he showed the images to Prum, who confirmed that they resembled melanosomes in every aspect.

To bolster the hypothesis that melanosomes can persist in fossils of extinct birds, Briggs wanted to find another example. He rummaged through the scientific literature for a good test case and found a description of a little Cretaceous feather from Brazil with preserved, distinct black and white color bands. Briggs thought that if we could show that this specimen also had preserved aligned melanosomes—but only in the dark bands because white coloration stems from a lack of pigment—we would have enough evidence to make our case. We managed to get the specimen on loan and put the entire block under the electron microscope. Lo and behold, when I examined the dark bands of this 108-million-year-old feather, thousands of little melanosomes aligned along the axes of the fine feather branches came into focus. When I looked at the white bands, in contrast, I saw nothing but rock matrix—which is exactly what one should expect in the absence of pigment.

FOSSIL of a *Psittacosaurus* preserves pigment patterns indicative of a type of camouflage called countershading.

Courtesy of Jakob Vinther and Bob Nicholls

PAINT BY NUMBER

SINCE THE PUBLICATION of our melanosome discoveries in 2008, my team and several others have described melanosomes and other pigments from additional fossils. Researchers have also investigated the chemistry of fossil melanin and substantiated our observations that melanin can survive for millions of years, almost chemically intact. Together with Caitlin Colleary, then a master's student at the University of Bristol in England, where I now work, we showed that the slight alterations evident in the fossil melanin are the result of sustained exposure to elevated pressure and heat in the ground. (A few investigators still maintain that the observed structures might be bacteria, but they are running out of options to support their claims.)

Some of our most spectacular findings have uncovered the colors of dinosaur feathers. In 2009 my Yale colleagues and I teamed up with Matthew Shawkey and Liliana D'Alba, both now at Ghent University in Belgium, and others to reconstruct the color pattern of *Anchiornis huxleyi,* a small, predatory, feathered dinosaur from China that lived around 155 million years ago. Like the Danish bird I had studied previously, the *Anchiornis* fossil had some dark stains visible to the naked eye, indicating the presence of organic material, probably melanin. But because we were aiming to reconstruct the pattern of its full plumage—a much more ambitious task than simply determining the presence or absence of melanosomes—we could not rely on these stains to tell us all we wanted to know. Instead we had to develop a way to objectively predict colors from the shapes of the melanosomes. To do this, we studied melanosomes from 12 black, 12 brown and 12 gray feathers of modern-day birds. By considering the length, width and aspect ratio of the melanosomes, as well as how much they vary in shape, we could predict feather color using a statistical method called quadratic discriminant analysis with 90 percent accuracy.

When we applied our method to the melanosomes of *Anchiornis,* the results were striking. Our statistical predictions indicated that the feathers that covered much of the creature's body were mostly gray. The long feathers on the animal's arms and legs, in contrast, were unpigmented by melanosomes and thus white, except for the melanosome-laden tips, which we predicted were black. (Modern birds often have black-tipped wing feathers. The melanin, in addition to coloring the feathers, also fortifies them against battering winds. Perhaps *Anchiornis* benefited from this strengthening property of melanin, too.) Most surprising, the feathers on the crown of the head contained impressions of round melanosomes—the "meatballs"—that would have given *Anchiornis* a ruddy crest. All told, this combination of colors made for a spectacularly flamboyant creature.

At around the same time we published our *Anchiornis* study, Fucheng Zhang, then at the Institute of Vertebrate Paleontology

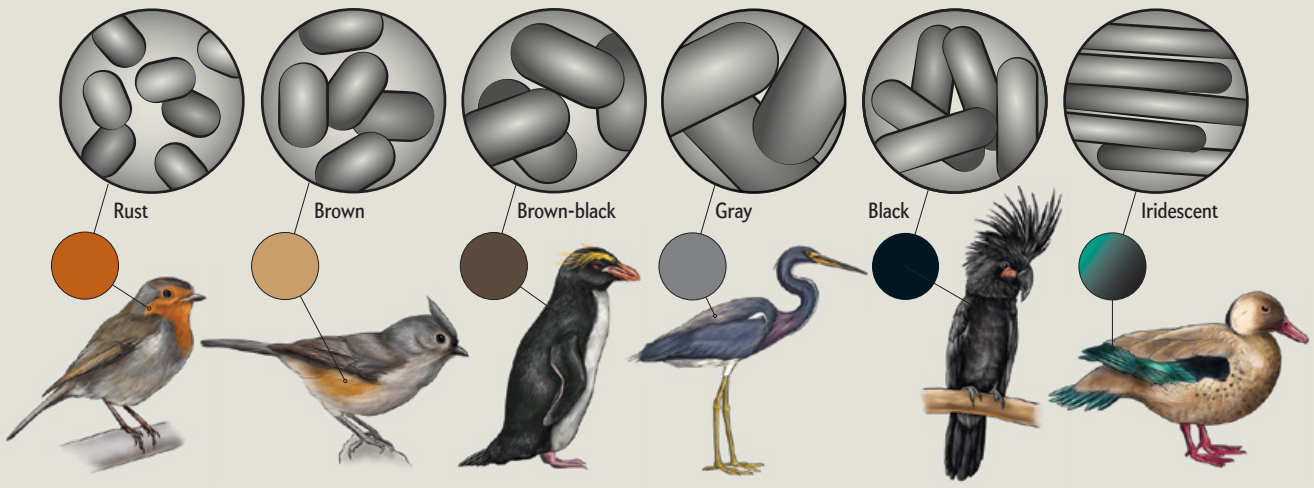
© 2023 Scientific American

In Living Color

Microscopic pigment-bearing cell structures known as melanosomes can persist in fossils for tens of millions of years. Studies of preserved pigments have allowed scientists to reconstruct the actual colors of a wide range of extinct animals, including a number of dinosaurs. These findings are not only revealing, for the first time, what these creatures really looked like, but they are also elucidating previously murky aspects of the animals' lives—from their activity cycles to the type of environment they inhabited.

Melanosome Size, Shape and Configuration Predict Color

Melanosomes contain two forms of melanin: eumelanin, which gives rise to black tones, and pheomelanin, which imparts rusty red hues. Combinations of these melanins and absence of pigment create brown, gray and white colors. Iridescence, for its part, stems from the stacking of melanosomes in ways that refract light. Analyses of melanosomes from feathers of modern-day birds have yielded a database that researchers can use to predict colors and patterns of extinct animals from the size, shape and arrangement of fossil melanosomes.



© 2023 Scientific American

Standing Out

Melanosomes preserved in a small dinosaur known as *Microraptor* reveal that this creature had showy, iridescent black plumage similar to a crow's. Paleontologists had suspected that *Microraptor* was nocturnal, based on the large size of its eye sockets. But modern birds with iridescent coloring tend to be active during the day, suggesting that *Microraptor* was actually diurnal.

Microraptor

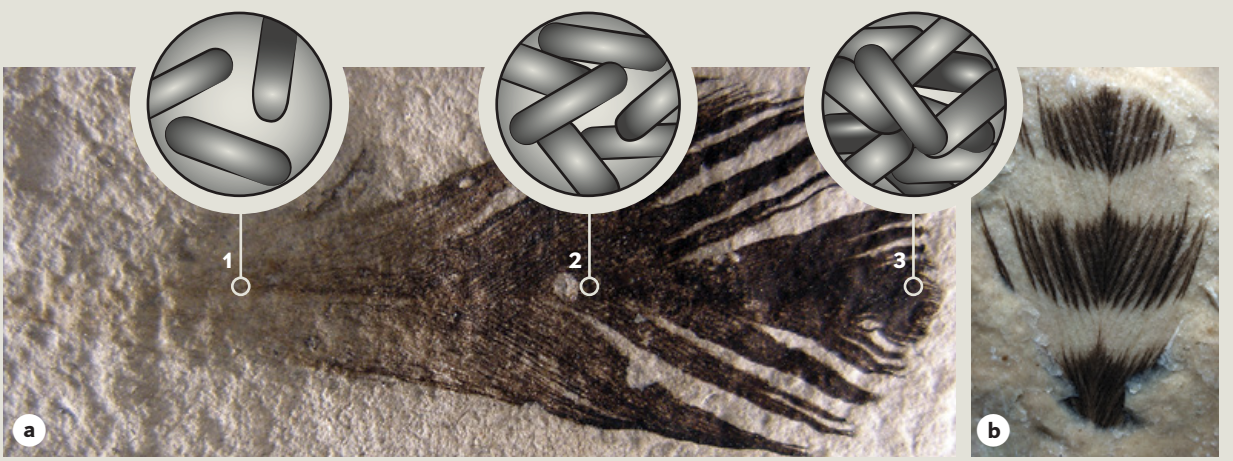
Psittacosaurus

Blending In

The melanosomes preserved in a *Psittacosaurus* fossil show that this animal had a dark back and light belly. This pattern, called countershading, is common in modern-day animals and helps to camouflage them from predators and prey. The specific form of countershading seen in *Psittacosaurus* suggests that the creature would have best blended into a habitat with diffuse sunlight such as that seen in a canopy forest.

Melanosome Density and Distribution Predict Pattern

Fossil feathers show how varying degrees of melanosome concentration can create patterns. For instance, the gradient pattern seen in a 55-million-year-old specimen from Denmark **a** arises from the combination of low melanosome concentrations that yield pale colors (*1*), intermediate concentrations that produce midrange tones (*2*) and high concentrations that form intense tones (*3*). In a 108-million-year-old fossil feather from Brazil **b**, dark and light stripes stem from melanosome-rich and melanosome-free areas, respectively.



Courtesy of Jakob Vinther *(feather photographs)*

Illustrations by Raúl Martín (dinosaurs) and Jillian Ditner (birds and melanosomes)

© 2023 Scientific American

and Paleoanthropology in Beijing, Michael J. Benton of the University of Bristol and their colleagues reported that they had found fossil melanosomes in a range of birds and dinosaurs recovered from 130-million-year-old rocks in China. The pattern of meatball melanosomes in one fuzz-covered dinosaur, *Sinosauropteryx,* implied that it had sported a reddish coat and a tiger-striped tail, making it the first known ginger dinosaur.

Since those early days our feather data set has grown to comprise hundreds of samples, including ones that allow us to accurately predict iridescence, the metallic sheen seen in the plumage of hummingbirds and peacocks, among other birds. Melanosomes responsible for this effect tend to be longer than typical ones, and they may even be hollow or flattened. The iridescence arises from the packing of the melanosomes within the feather. Certain configurations refract light in ways that create different colors, depending on the angle at which the animal is viewed or illuminated.

Amazingly, in 2009 we found evidence of iridescence in a 49-million-year-old fossil feather from Messel, Germany. The fossil, kept at the Senckenberg Museum in Frankfurt, preserves the original arrangement of melanosomes that generated the iridescence. They were packed into a dense, smooth layer found in the finest branches of the feather fossil, the barbules. There the melanosomes occurred strictly on the farthest edge of the feather and on the top surface, the only part that was not obscured by other, overlapping feathers. We deduced that the tips were iridescent because that arrangement of melanosomes is known to produce what is called thin-film interference, the kind that occurs when gasoline floats on water and creates a vivid rainbow of colors.

It was not long before we discovered evidence of iridescence in an actual dinosaur—a crow-size creature from China with wings on all four limbs. Dubbed *Microraptor,* it was a primitive cousin to *Jurassic Park*'s *Velociraptor.* The movie depicted *Velociraptor* with scaly skin, but scientists now know that both these dinosaurs were, in fact, covered in feathers. In *Microraptor,* the preserved feathers contain long, sausage-shaped melanosomes arranged to bend light in eye-catching ways. Its plumage thus would have been black, with the same shiny sheen as a crow's. *Microraptor* is not the only extinct creature now known to have had that rainbow shimmer. Jennifer Peteya of Oberlin College and Ghent's Shawkey have described shimmering iridescence in an enantiornithine bird, called *Bohaiornis,* and a Jurassic theropod with a big, fan-shaped tail, named *Caihong.*

THE PATTERNS WERE SUBTLE, WITH FINE VEINING, DOTS AND STRIPES.

MORE THAN SKIN DEEP

BEYOND ALLOWING PALEONTOLOGISTS and artists to reconstruct extinct organisms more accurately, fossil pigments are revealing previously unknown facets of the daily lives of both dinosaurs and other long-gone creatures. For instance, experts had presumed that *Microraptor* was nocturnal, based on the large size of its eye sockets. But our discovery that it possessed iridescent plumage suggests otherwise because in modern birds such coloration is typically found in species that are active in the daytime. The bold coloring of *Anchiornis,* for its part, probably helped to attract mates or served as some other kind of display, as occurs in flashily dressed modern birds. Thus, color patterns may provide a way to test behavioral hypotheses about a species using a different line of evidence than usual.

Fossil pigments in one species can also illuminate aspects of the other species with which it interacted. Among insects, most color patterns evolved not to help the creatures attract mates but rather as a tactic to avoid getting eaten. Their pigments can thus provide clues to their predators. Fossils of insects called lacewings offer a fascinating example. Between 170 million and 150 million years ago certain distinctive color patterns made their evolutionary debut in insects. Perhaps the most dramatic pattern to emerge during this time was the eyespot, a marking that resembles the eye of a different kind of animal and serves to startle predators approaching their prey at speed from a distance. Lacewings are among the first creatures known to have had eyespots. What kind of predator were they defending themselves against? Most color patterns of modern insects have evolved as a defense against birds, which are their main predators nowadays. But the lacewings' eyespots predate the origin of birds as we know them. Their predators were instead most likely a small group of dinosaurs called the paravians, which are known to have lived at the same time as these lacewings and are thought to have given rise to birds. Although the fossil record of paravians themselves has not allowed us to unequivocally pinpoint when flight evolved in this group, the appearance of these eyespots in the lacewings hints that some paravian dinosaurs had taken wing by this point and were exerting birdlike predation pressure on the insects.

Other fossil melanosome discoveries have enabled my collaborators and me to reverse engineer the environment in which extinct organisms lived. Our first foray into this realm of investigation began with a particularly splendid fossil of a small, plant-eating dinosaur called *Psittacosaurus,* a relative of *Triceratops*. These dinosaurs' skeletons are quite common in northeastern China and are often complete. This specimen stood out even in that good company, however. A thin film drapes its body—the remains of the skin, including delicate scales. And its tail displays long, filamentous bristles that may be precursors to feathers. Previous discoveries of dinosaur feathers have all come from the mostly carnivorous theropod group of dinosaurs. The bristles on *Psittacosaurus,* a distantly related member of the plant-eating ceratopsian group, hint that plumage might have been far more widespread among the dinosaurs than previously thought.

When I first encountered the specimen in 2009, a year after we had announced the discovery of melanosomes in fossil birds, I saw right away that it preserved evidence of beautiful color patterns all over the body. The patterns were subtle, with fine veining, dots and stripes. And I could see that the animal had a dark back that gave way to a pale belly. That kind of dark-to-light color gradient from back to belly counteracts the light-to-dark gradient created by illumination from the sun. This pattern, known as countershading, is common among modern animals ranging from dolphins to deer, helping both predators and prey blend in with their surroundings and thereby elude detection.

© 2023 Scientific American

I eventually showed the *Psittacosaurus* pattern to Innes Cuthill, who is part of a group that studies camouflage at the University of Bristol. It was then that we realized we had the opportunity not only to study countershading in a dinosaur but also to deduce from the fossil alone what kind of environment the creature lived in. To reconstruct an animal's habitat, scientists usually gather clues from fossils of other animals and plants found nearby. This kind of approach is problematic, however, because oftentimes the site where a fossil is discovered is not where the organism lived. The Chinese psittacosaur, for example, was recovered from sediments of an ancient lake. The creature was clearly not aquatic, so its remains must have been transported to the lake from the surrounding terrestrial environment, perhaps by moving water. Our study might be able to provide clues about that setting—specifically, the light conditions under which this dinosaur evolved its camouflage.

Cuthill and his collaborators had recently studied countershading in modern ungulates, the group that includes horses, antelope, camels, pigs and rhinoceroses. Although countershading by definition involves darker coloration on the back and lighter coloration on the underside (except in some animals, such as caterpillars, that live their lives upside down), the intensity of those shades and the nature of the transition from dark to pale differ from species to species. Cuthill's team wanted to investigate how well that variation correlates to variation in the lighting conditions found in different environments. Because sunlight varies depending on the latitude at which an animal lives, as well as the density of vegetation in its habitat, the researchers had theorized that ungulate countershading, too, should differ according to latitude and habitat. Their findings bore out that notion. Broadly speaking, if an animal lives in open habitats, the direct sunlight will create a shadow high on the body, with a very sharp transition to the illuminated areas. These animals usually exhibit countershading that matches this pattern, with dark backs that almost immediately give way to pale bellies and little intermediate coloration in between. Pronghorn antelope offer a great example of this kind of countershading. In closed habitats, in contrast, the diffuse light that filters down through the vegetation scatters in all angles, producing a shadow that hangs farther down the body and transitions to the illuminated area gradually. White- and black-tailed deer, common in North American forestlands, exhibit this pattern.

We knew from our visual inspection of the *Psittacosaurus* fossil that it had countershading of some sort. We therefore carefully projected the pigment pattern onto an accurate, life-size model of the dinosaur, which we accomplished by enlisting the help of British paleoartist Bob Nicholls. Through this work we determined that the transition from dark to light occurred low on the belly and tail in *Psittacosaurus.*

To test the function of the dinosaur's color pattern, we painted a second copy of the full-scale model gray. Next we photographed this model in a range of daylight conditions, from gloriously sunny to oppressively cloudy, as well as in open land and underneath conifer trees to capture the shadows cast on it. Next we inverted the dark and light shades in the photographs, effectively creating the ideal countershading patterns for concealing the animal in each of the lighting conditions. By comparing our reconstruction of the actual countershading pattern of the *Psittacosaurus* with the idealized countershading patterns, we determined that the animal's coloring would have best camouflaged it in a habitat with diffuse light, such as that seen in a canopy forest.

Our work on *Psittacosaurus* didn't end there. In 2021 we reconstructed its cloaca—the multipurpose orifice for defecation and breeding—in great detail. We could show that while the rest of *Psittacosaurus*'s body was camouflaged, the flaring lips defining the cloacal opening were heavily pigmented and must have been used for signaling of some sort, most likely for mating.

Subsequent studies on countershading found that the tiger-striped *Sinosauropteryx* was adapted to live in open environments with bright sunshine from above and that it carried a striped tail and a bandit mask of pigmented feathers over its eyes. *Psittacosaurus* and *Sinosauropteryx* are found in the same fossil beds, but their countershading tells us that they came from different environments.

A nearly six-meter-long ankylosaur, *Borealopelta,* was found in the approximately 112-million-year-old marine oil sand beds in northern Alberta, Canada. It was colored with reddish-brown pheomelanin and was countershaded. Living land animals of such size are not countershaded, because there are no predators big enough to threaten them. In other words, for such a big creature to maintain its countershading from generation to generation, the Cretaceous predators must have been vicious enough to have threatened it. Perhaps that doesn't come as a big surprise considering the spine-chilling Herculean dimensions of the theropod predators back then. But now we have rock-solid proof in fossil color pattern of their gruesomeness.

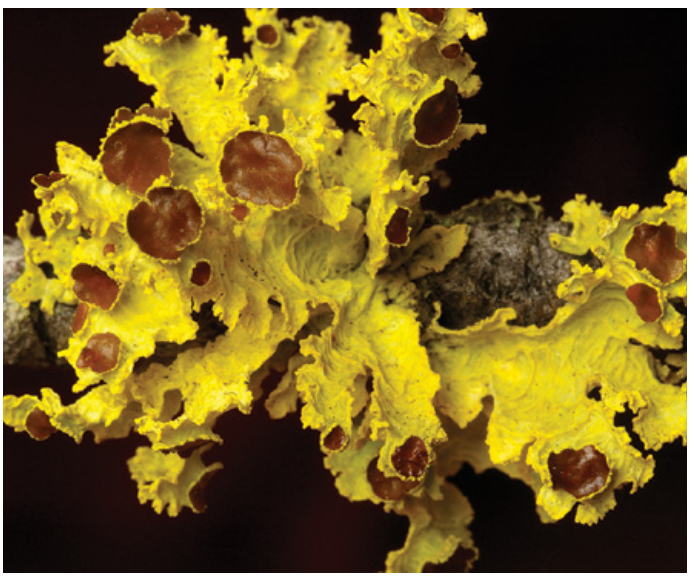
A VIVID FUTURE

SCIENTISTS STILL HAVE MUCH TO LEARN about paleocolor. Our ability to see broad categories of color in fossils—those that stem from the shape and arrangement of melanosomes—is already a massive leap forward from what we knew about ancient hues 15 years ago. But there are other pigments to look for in fossils, including carotenoids, which produce bright reds and yellows, and porphyrins, which produce such hues as green, red and blue. These pigments have turned up in the fossil record on occasion. Researchers have identified carotenoid pigments derived from fossil bacteria dating back several billion years; porphyrins are preserved in a blood-engorged mosquito from 46 million years ago and in the eggs of a 66-million-year-old dinosaur known as an oviraptorosaur. Pigments not known from modern organisms have come to light, too, including some from fossil sea lilies and algae dating to between 300 million and 150 million years ago.

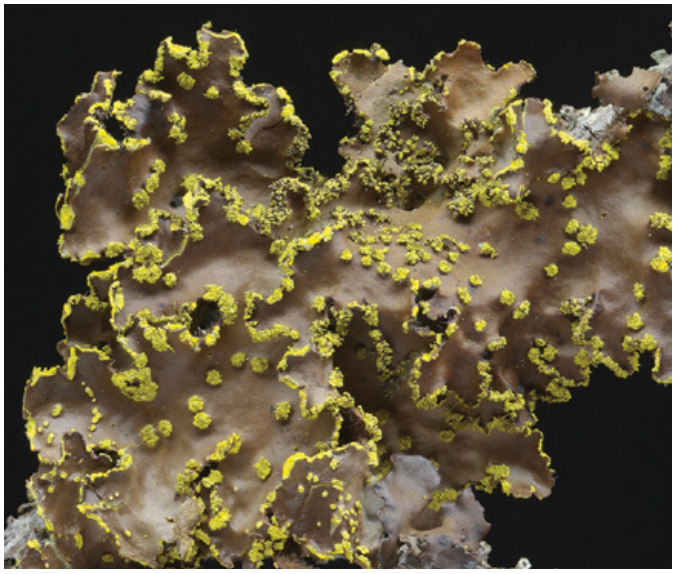
We will probably encounter limitations to the detail with which we can reconstruct paleocolors; over millions of years some information is bound to be lost forever. In addition, because exceptional fossils with organic preservation are rare and precious, we must restrict destructive chemical sampling of them. As techniques advance, however, the new discoveries they afford will undoubtedly change our understanding of the past faster than ever before. Each one will bring us that much closer to seeing dinosaurs and other prehistoric creatures as they really were, in full Technicolor glory. SA

Jakob Vinther thought he was going to become a botanist until he found his first fossils at age 11 while attending summer camp in his home country of Denmark. Today he is an associate professor of macroevolution at the University of Bristol in England. His research focuses on pigments and other molecules preserved in the fossil record.

© 2023 Scientific American



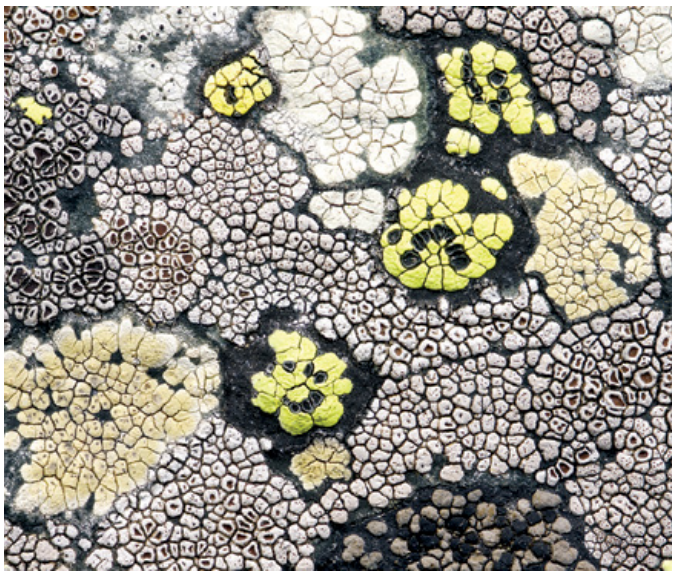
© 2023 Scientific American

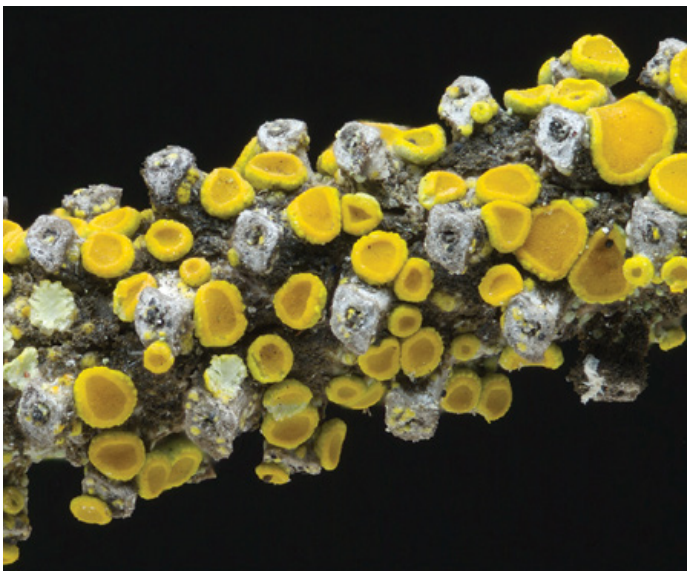


The Meaning of Lichen

How a naturalist's observations in the wilds of British Columbia inspired a scientist to discover hidden symbioses—overturning 150 years of accepted scientific wisdom

By Erica Gies





© 2023 Scientific American

TREVOR GOWARD LETS ME LEAD, SO WE TRAVEL THROUGH THE MIXED FOREST at my pace. This is a nod to his rangy 6'5" figure and the rapid strides he makes across barely discernible deer and bear paths on his land adjacent to Wells Gray Provincial Park in British Columbia. But mostly he is making space for my observations, my innate way of experiencing the landscape. What engages me? How do I see? I brake in front of a small, white-green growth on the trunk of an aspen. "Look there!" he says excitedly, inviting me to peer into the ragbag lichen through a magnifying lens. Suddenly I am in another world, looking down whorls studded with black dots into little caverns sprinkled with superfine dust.

Goward, white hair sticking up haphazardly, is wearing three flannel shirts on this crisp fall day. A hand lens hangs on a string around his neck, and an Australian shepherd named Purple trots along with him. He seems more mountain man than scientist, a naturalist in the tradition of Charles Darwin or Henry David Thoreau. Goward's scientific love is lichens—those growths that look like little mosses or colored crusts stuck to trees and rocks everywhere. He is inseparable from this place, where he has spent most of his adult life after growing up in a city south of the park. During my visit he tells me that now, in his 60s, he rarely leaves. "It has become my center of spiritual gravity," he tells me.

It's not hard to see why. Most of the park has no road access and is rarely seen by humans. Wells Gray's 1.3 million acres were formed by volcanoes and glaciers; its river valleys, sheer rock mountains, alpine meadows and waterfall spray zones foster rich biodiversity. "I came to understand that the lichens here are pretty special," among the world's most diverse, Goward beams. There are hundreds of species and counting. His careful attention to this one place, like conservationist Aldo Leopold's to his beloved Sauk County, Wisconsin, allows him to see connections that others might miss.

Goward stumbled on lichenology when he was educating himself about different branches of nature. "I made a point each year of learning as much as I could about a different taxonomic group. One year it was birds, then plants, then mushrooms, then insects." When he got to lichens, he was smitten. Since then, despite being self-taught, he has become the go-to expert in central British Columbia for everyone from atmospheric scientists to gold prospectors to caribou biologists. Several lichen species have been named after him. He has published three taxonomic guides to lichens and has earned a spot as an associate member of the University of British Columbia's botany department.

And yet Goward is a maverick in the scientific world. His radical thought experiments about lichens, published in provocative essays available on his website, Ways of Enlichenment, have been both ridiculed and lauded—but largely ignored by most researchers because he holds no scientific degrees and because many of his ideas are not supported by rigorous data. Still, Goward's astute observations and deep thinking follow in the footsteps of Darwin's and Thoreau's approaches—which, much more than laboratory science, formed the basis of the theories of evolution and ecology. People who are open to considering his ideas say they come away with mind-expanding food for thought about lichens, biology and all life. Goward's longtime friend and sometimes co-author Toby Spribille, a lichenolo-

Tim Wheeler *(preceding pages)*

© 2023 Scientific American

NATURALIST Trevor Goward examines a ragbag lichen on an aspen. His deep observations have upset biology, like those of his predecessors Charles Darwin and Henry David Thoreau.

gist at the University of Alberta, says Goward's essays contain many gold nuggets: "Frankly, I think that they are brilliant."

In the forest, Goward exudes a quiet, ebullient joy, the yang to his yin: a dark, realistic assessment of humanity's folly. When we stop, he leans on a sturdy walking stick and delivers extended soliloquies about how elements of the ecosystem interact. Reading the lichens informs him about soil chemistry, rain patterns and plant nutrients. He shows me a species growing on a hemlock, unusual because conifer bark is generally too acidic to support these kinds of lichens. So why are they there? In a 2000 paper Goward and his co-author André Arsenault found that the answer lies in a mature trembling aspen nearby. Water dripping from its branches becomes a leachate, which, when it falls onto the conifer's bark, lessens the acidity, allowing the lichen to thrive. They dubbed this interaction the drip-zone effect.

Goward learns from every life-form, including Purple, who waits on us patiently when she is not conducting her own observations: Scat from a pine marten. Red squirrel chatter! Although Goward knows French, Latin, and some conversational German and Swedish, he remarks that "mostly these days I speak lichen and maybe a bit of dog." He says he can learn from Purple's way of seeing. That may seem eccentric, but Goward respects First Nations peoples' ways of knowing, and learning from animals is a storied human tradition.

Modern science tends to ignore outsiders. But reductionist science is not the only way of knowing things. Naturalists were the forebears of science. Humans once lived much closer to the land and were keen observers who had a profound understanding of nature's interactions. Today biology tends to be focused on molecules, and failure to look up from instruments in the lab to actually observe how pieces interact in the natural world sometimes undermines discovery. A clinical focus can lead scientists to miss big-picture connections, such as an emerging understanding that networks may be a more enduring life-form than individuals. Indeed, it was an idea from Goward that inspired Spribille's lab work while he was a postdoc at the University of Montana. That work paid off with a major advance: a July 2016 cover story in *Science* that rocked the ossified world of lichenology. The discovery called into question the very nature of the lichen symbiosis, shedding fresh light on how symbioses across biology work, how natural selection proceeds and even how to define life-forms.

Lichens are both ubiquitous and fascinating. Perhaps more than 500 million years old, they occur on every continent and can thrive in some of the most inhospitable places on Earth. They even survived for a year and a half

Erica Gies

© 2023 Scientific American

CLOSE EXAMINATION has revealed that *Bryoria fremontii* is not simply a partnership between a fungus and an alga, as long thought; a yeast is also involved.

Tim Wheeler (*this page and opposite page*)

in space, fully exposed to cosmic radiation, ultraviolet irradiation and vacuum conditions. The approximately 14,000 species of lichen come in a variety of forms: flat rounds on stones, scalloped leaves nestled among mosses, crusts clinging to tree bark, flowing strands hanging from branches, tiny trumpets tipped in red.

For centuries people thought lichens were plants (and then fungi). Then, in the 1860s, Swiss botanist Simon Schwendener discovered that they were a partnership between a fungus (an organism classified in its own kingdom because, unlike plants, it cannot make its own food) and an alga, an organism that feeds itself with photosynthesis but lacks the roots and stems of plants. The fungus apparently provided the structure of the lichen, and the alga provided food for the fungus via photosynthesis. (Later it was discovered that in some lichens, a cyanobacterium provided the food—and a handful of species contained both an alga and a cyanobacterium, along with the fungus.) Schwendener's discovery, at first resisted by the scientific community, ultimately made lichens the poster children for symbiosis, a mutually beneficial interaction among organisms. Since then, science has found symbioses across nature, including among the trillions of nonhuman microbes that cling to the scaffold of our bodies.

Science over the past two centuries has largely viewed molecules, cells and species as individuals. Symbiosis challenges that notion. "Within a lichen," Spribille says, "algal cells and fungal cells may experience each other as individuals, but together they form a lichen that the feeding caribou sees as an individual: tasty." Natural selection happens on both scales simultaneously. Just as light is both a wave and a particle, the fungus and alga are both individuals and parts of a whole. Science's reductionist focus has made it nearly impossible to fully understand symbiosis, Spribille says. "Ecology was supposed to be the science of natural process and synthesis, but its backbone is severely strained under the mathematics of individuality."

In July 2016 Spribille and his co-authors took a major step forward in that understanding. Their big reveal in *Science:* many lichens have a second fungus in the house.

At the heart of their study is a pair of lichens to which Goward had drawn Spribille's attention: *Bryoria fremontii,* which is hairlike and often brown and eaten by northwestern Indigenous peoples, and a similar lichen, *Bryoria tortuosa,* which is often a yellowish green and is toxic, with high levels of vulpinic acid. The two posed a fascinating conundrum. Despite their differences, a genetic analysis published in 2009 by Saara Velmala of the University of Helsinki and her colleagues, on which Goward was a co-author, showed that the two species consisted of the same fungus and same alga. Spribille recalled how this perplexing finding infected them both. "[Goward] took the question of how could these two different lichens—one of which is toxic, for God's sake—be identical." The question would not let go of Goward. And when Goward wrote about it, "by extension, it wouldn't let go of me."

Aside from their different appearances and levels of vulpinic acid, Goward observed that the two lichens also had slightly different ecologies. Although they grew in some of the same places, *B. tortuosa* was found only on the summer-dry fringes of *B. fremontii*'s larger territory. In 2009 he proposed that lichens are formed not by the shape of their fungal partner but by a series of decisions made during the developmental dance between fungus and alga. One lichen can look different from another that is composed of the same partners because it took different turns during development. Goward suggested that the difference between the two species of *Bryoria* might stem from each of them having a different relationship with a third life-form, a bacterium.

After five years of work in the lab, Spribille and his colleagues discovered that both *Bryoria* species did include a third partner. But it was not a bacterium; it was another fungus, known as a basidiomycete yeast. The toxic *Bryoria* contained a lot more of the yeast than the edible one. The researchers also demonstrated that the yeast was not a contaminant but had evolved with the other partners for more than 200 million years. Broadening their search to lichens across the globe, they found the yeast in 52 other sets (genera) of lichen. The finding dramatically expanded the world's understanding of lichens, opening the door to other insights. "Only now are we beginning to see that lichens have pulled off a rare feat in evolution: a large multicellular organism but built entirely of microbes—and here's the amazing thing—without a scaffold," Spribille says. "Self-assembling, self-replicating, generation after symbiotic generation."

Goward first became interested in *B. fremontii* and *B. tortuosa* when he read ethnobotanist Nancy Turner's

© 2023 Scientific American

1977 paper about *B. fremontii*'s importance to First Nations peoples. She said that women elders could easily distinguish the edible from the nonedible lichens. Although the two can have different coloring and a slightly different shape, they can also look quite alike. Elders use clues such as location, color and the types of neighboring lichens to tell them apart. When Stuart Crawford, a friend of Goward's with a degree in ethnobotany, showed bundles of the two lichens to an elder and conservationist from the Neskonlith band, the late Mary Thomas, she correctly identified the edible one every time.

Local people's wisdom does not always jibe with scientific explanations, Crawford says, but the result, based on observation, is correct. The locals told Crawford that they wait for *B. fremontii* to "ripen" on the tree. In fact, lichens do not ripen as do fruits and vegetables, but the darker color and its growth pattern on trees help the people distinguish it from its poisonous twin. These other realms of knowledge about *Bryoria* would have added interesting context to the *Science* paper, Spribille says, but "it didn't fit the word limit."

BRYORIA TORTUOSA hosts the same fungus and alga as the edible *B. fremontii,* but it has a much higher concentration of yeast, and it is poisonous.

Three months after the paper was published, Crawford, who knows Spribille through Goward, got around to telling him something amazing. For years Crawford had been collecting writings from around the world—ancient Egypt, modern Mexico, medieval Russia, the Biblical Middle East, a European cookbook from the 1950s—of people using lichens to make bread and alcoholic beverages. In some cases, they were using them explicitly for leavening and fermentation. On some level, Crawford realized, people knew that lichens contained yeast or functioned like yeast. When he was working on his master's degree in Victoria, B.C., Crawford discussed the notion with a local microbrewer, who told him, "If you can figure out the recipe, I'll do a batch of beer" with it.

Spribille's openness to Goward's unconventional ways of thinking is perhaps a reflection of his hard-won path to science. He grew up in a fundamentalist Christian family in northwestern Montana, where his parents pulled him out of school after fourth grade to protect him from "the influences of the world." Spribille is telling me this via Skype from Austria, where it is late at night and his wife and young daughter are sleeping. His rectangular glasses frame blue eyes that frequently squeeze shut while he is talking, as if communicating with me is a little painful.

Circumstance could not restrain Spribille's intellectual curiosity. Intrigued about organisms he saw in the wild, he sought answers from biologists at a local U.S. Forest Service outpost. Eventually they recommended him for a job surveying vascular plants, and he could call up professors and authors with his burning questions. Goward was on Spribille's call list. "Trevor kept me on for two and a half hours," Spribille says fondly. That was decades ago. They have co-written several papers, and "we still haven't run out of things to talk about." Early on, Goward told Spribille that he had ideas that would turn lichenology upside down. "He said that I was delusional," Goward recalls. "But he wanted to hear the ideas."

Ultimately Spribille felt a keen desire for a formal education. He took the high school equivalency examination and found an opportunity to go to college in Germany. He later earned a Ph.D. in lichenology at the University of Graz in Austria and in 2017 joined the faculty at the University of Alberta, where he studies the ecology and evolution of symbiosis. During his postdoc at the University of Montana, he met John McCutcheon, one of the co-authors on the *Science* paper and head of the lab in which the work was done (McCutcheon has since moved to Arizona State University). McCutcheon credits the breakthrough to technological advances that allowed the researchers to find the tiny yeast and to cooperation among diverse co-workers. But also critical, he says, was Spribille's ability to look beyond what was assumed to be true. The human mind's tendency to restrict itself is part of what kept this yeast hidden for so long, he says: "When you're used to thinking there's just one fungus there, that's what you see."

Spribille, in turn, credits Goward with having "a huge influence on my thinking. [His essays] gave me license to think about lichens in a way that was not orthodox and freed me to see the patterns I worked out in *Bryoria* with my co-authors." Yet even with that, Spribille says, "one of the most difficult things was allowing myself to have an open mind to the idea that 150 years of literature may have entirely missed the theoretical possibility that there would be more than one fungal partner in the lichen symbiosis."

Although he appreciates his education, Spribille maintains that academia's emphasis on the canon of

© 2023 Scientific American

what others have established as important is inherently limiting. "You have this culture of prepared minds that makes it extremely difficult to think outside the box," Spribille says. "It creates the box."

That sounds plausible to Jonathan Foley, executive director of Project Drawdown, who has a Ph.D. and had an acclaimed academic career. (Foley serves on *Scientific American*'s board of advisers.) When it comes to ideas, "the ivory tower is now an ivory fortress," he says. Academic culture's incentives to publish in accepted journals, get funding and obtain tenure are "not aligned with being wildly creative." After Sputnik, science became hyperprofessionalized, Foley says—"kind of Science, Inc. I think we lost part of our souls." The extreme specialization required for the biotech-heavy, molecule-focused world of biology today eliminates time to study taxonomy or epistemology. "There are people getting degrees in biological sciences at the best universities in America today who don't know the names of anything outdoors, who have never studied anything larger than a cell," Foley notes. That means a lot of biologists are lab-bound and rely on people like Goward to find the species they would like to study or even to suggest ideas for studies.

LICHENS GROW on every continent, and they survived a year and a half on the outside of the International Space Station, fully exposed to cosmic radiation.

Also worrisome to Spribille is that his own students are petrified of being wrong, a psychological state incompatible with breakthroughs. For a counterexample, he points to Goward. In the case of *Bryoria,* Goward surmised that a third partner was present, although he incorrectly thought it was a bacterium. But being correct "is not the criterion of a brilliant mind," Spribille says. On the contrary, brilliant minds are characterized by indefatigable curiosity and questioning, traits that Spribille tries to encourage in his students. "I tell them, 'Just put all the ideas out there. Nobody here is going to make you feel bad about throwing out an idea that we may then not use.' I live by that."

Some of the most serious problems science is trying to solve today—climate change, loss of biodiversity, food and water security—require big, integrated views from multiple perspectives. Stepping out of the lab and back into nature to observe how natural systems actually work is a critical first move. One biologist at the University of the South challenged himself to try it. David George Haskell spent a year sitting in a square yard of old-growth forest in Tennessee, just observing, and wrote a Pulitzer-nominated book about it, *The Forest Unseen*. The experience was profoundly humbling, he says. "You wake up to the extent of your own ignorance. I'd been through decades of training and teaching as a biologist and had published scientific papers and so on, and sitting down in the woods, I realized I know so, so little about this place." From that humility sprouted seeds of curiosity and dozens of questions about relationships among plants and animals, their ecological history, and how that related to climate and geology. Haskell eventually became an adviser to the New York City–based Open Space Institute, helping it to identify lands for conservation that are most likely to be climate-resilient.

If knowledge comes mostly through reading scientific literature, "we're several steps removed from the actual phenomena we're discovering," Haskell explains. And while instruments are important for helping scientists understand the world, "our bodies come preinstalled with all these amazing apps, and they connect directly into our consciousness," he says. "Through literally coming back to our senses, we can learn so much about the world."

Goward has turned this ethic into a way of life. His house, named Edgewood Blue, on 10 acres abutting Wells Gray, has running water for a shower and sinks but no toilet. As I put on my coat and shoes one evening to head to the outhouse, Goward's partner, Curtis Björk, a botanist, encourages me to look up and admire the Milky Way, vivid in the lack of light pollution. When I ask Goward why he and Björk have no toilet, he says they appreciate being forced to go outside every day, even in the depths of winter. On trips to the loo he has seen the Northern Lights and passing moose. When I jokingly whimper about getting wet or cold, or chomped by summer mosquitoes, or even stalked by the cougar that recently swiped a neighbor's pigs, Goward is unapologetic: "That's real. Life isn't always comfortable."

To Goward, the real danger lies in separating ourselves from the natural world, living ensconced in cities, ignorant of how badly we are degrading nature. Haskell agrees and points out that this separation has ethical implications. "Trees, fungi, salamanders … these are our blood kin, if you believe Darwin." When we do not know the world, we have an imperfect sense of right and wrong in how our own behavior impacts ecosystems, he says.

Erica Gies (*this page and opposite page*)

© 2023 Scientific American

But for scientists who may not have a year, or 30, to spend contemplating the wild, collaborations or friendships with people outside the academy or from different disciplines can open space for new discoveries, as they did for Spribille.

The decor at Edgewood is dominated by books, which serve as de facto wallpaper, lining homemade shelves in most rooms. The kitchen is Björk's fiefdom, and when dinner is ready, he sends Purple to fetch the humans. Purple eats at the table with us, displaying excellent manners.

Also served at the table are wide-ranging discussions. "We try to make this a place where anyone can express their ideas," Goward pronounces. The strengths and failings of modern science are a frequent theme, with conversation on the topic driven primarily by abiding affection for it. Life and human relationships also take the spotlight, sometimes with quirky analogies to lichens. Although Goward is confident—sometimes bordering on arrogant—about his own ideas, he is eager to consider new information. His dialogue is peppered with references to authors. I mentioned Haskell's book and an essay by the late Ursula K. Le Guin, and he had read them by the next time we talked.

AT HIS WOODLAND HOME, Goward hosts an ongoing parade of biologists, poets and astrophysicists who explore ideas about exotic life-forms and the pitfalls of humans separating from the natural world.

With the scientific world often reluctant to publish him, Goward spreads his ideas one person at a time. He and Björk host an ongoing parade of biologists, aspiring naturalists, poets, geographers, ecologists, astrophysicists and journalists who stay for a day or a week or longer in return for doing a bit of work at Edgewood. Both Spribille and Crawford have become regular guests. "I've done a lot of landscaping on his place," Crawford says proudly. "We have great and intellectually stimulating conversations."

Goward would like to create a more formal venue for learning, to increase "biological literacy" in the next generation, and has donated half of his land as a research center to Thompson Rivers University in nearby Kamloops, where he grew up. He also periodically invites people from various disciplines to meet for a few days of discourse.

On my visit Goward delves into one of his pet lines of inquiry: What are lichens, really? Are they organisms? Fungal greenhouses? Algal farmsteads? Ecosystems? Networks?

What you think lichens are might depend on your perspective. Lichens have the scientific names of their fungi, which can create an implicit bias toward seeing the fungus as in charge, a limited perspective that Goward admits to having once upon a time. Today he sees lichens as a kind of koan. "The lichen by its very nature exists at a portal, a doorway," he says. "If you look in one direction, it's an organism. If you look in the other direction, it's an ecosystem." Goward's essays argue for seeing lichens not as their fungal or algal parts or even as ecosystems or organisms. Rather they are all these things, biological systems encapsulated in a membrane: lichens as emergent property. After all, the lichens that were sent into space survived when their algae alone did not.

Thinking of lichens as systems fits with a larger shift in biology from viewing the individual as the fundamental unit of life to highlighting community or partnerships. "Whether it is the microbiome within the human body, or trees interacting with fungal partners belowground, or lichens ... we're seeing that networked relationships are more fundamental and persist longer within biological systems than individuals do," Haskell says.

To Goward, lichens are the organisms that are most obviously about relationships. As such, they provide insights into all of life. "Lichens are my window," he says, "but it's the act of looking at the world that's the interesting thing." Systems hold together in the long term only if the parts consider themselves integral to the whole and if the whole protects the parts, as lichens do. "That's what's going wrong with us," he says. "As individuals, we're not concerned with the whole." SA

Erica Gies writes about science and the environment from Victoria, B.C., and San Francisco. Her work appears in the *New York Times*, the *Guardian*, the *Economist*, *Ensia*, *National Geographic*, and elsewhere.

© 2023 Scientific American

The Power of Viking Women

Analyses of ancient North Atlantic textiles show that Viking and medieval women wielded considerable cultural and economic influence

By Francine Russo

Illustration by Chase Stone

© 2023 Scientific American

© 2023 Scientific American

ARCHAEOLOGY HAS A REPRESENTATION PROBLEM. FOR most of the time that scholars have been probing the human past, they have focused mainly on the activities of men to the exclusion of women. There are a couple of reasons for this bias. One is that the kinds of artifacts that tend to be preserved well are made of inorganic materials such as stone or metal, and many are associated with behaviors stereotypically linked to men, such as hunting. Another reason is that early archaeologists were mostly men and more interested in men's work than in women's. As a result, our understanding of past cultures is woefully incomplete.

In recent years archaeologists have sought to fill that gap in our knowledge, in part by taking a closer look at traditionally ignored remains such as textiles, which had long been dismissed as trivial. Cloth rarely survives the centuries because it decomposes easily except under ideal preservation conditions. But even in a fragmentary state, it contains a wealth of information about the people who made and used it.

Michèle Hayeur Smith, an anthropological archaeologist at Brown University, has been at the forefront of efforts to glean insights from ancient cloth, scouring archaeological sites and museum collections for textiles that could illuminate the lives of women in early North Atlantic societies. Her work has shown that the Vikings never would have expanded their known world without the women's work of weaving.

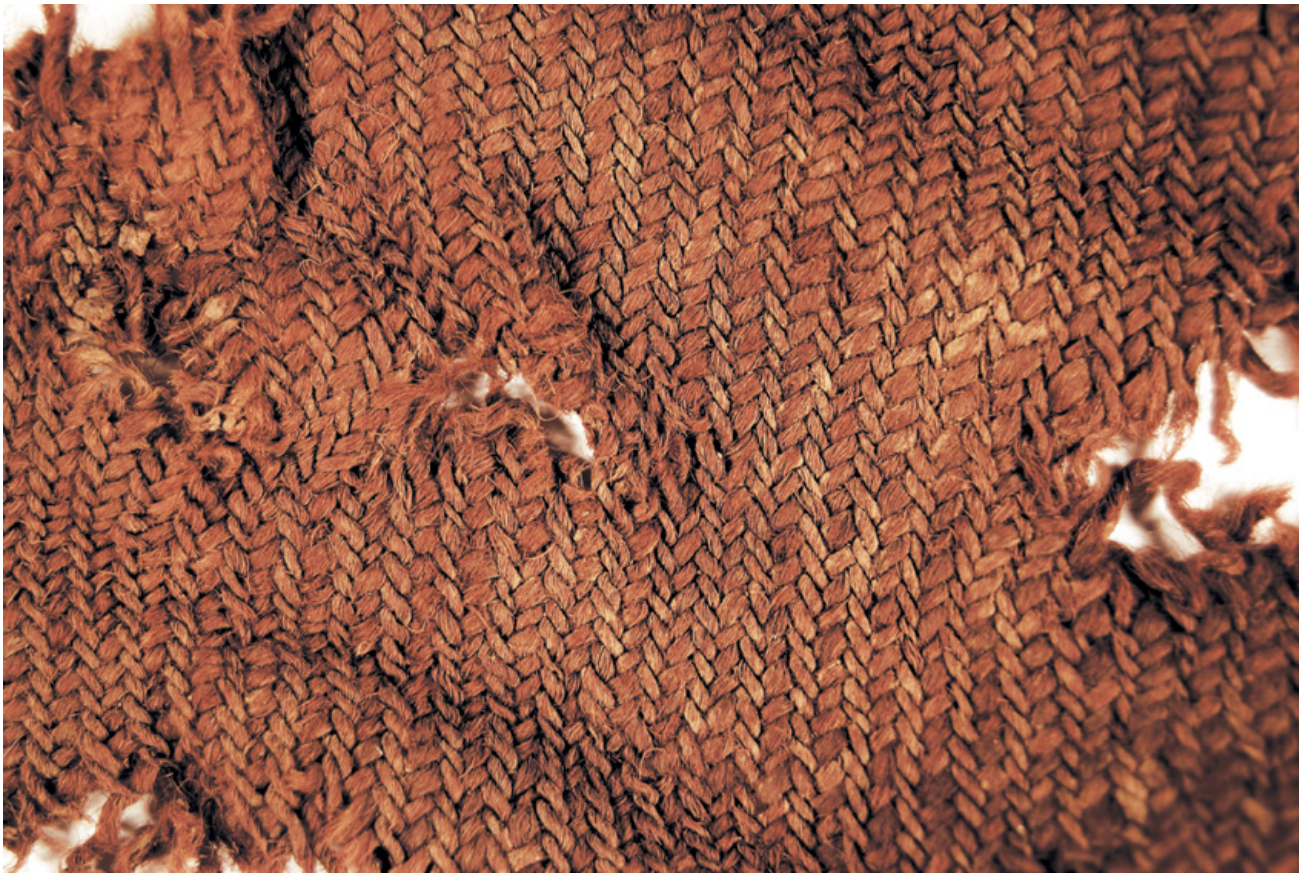
Hayeur Smith's study of early North Atlantic textiles took off from the basement storage area of the National Museum of Iceland, its rows of metal shelving bursting with boxes and bags of dirt-covered cloth. She first visited in 2009 to inspect the museum's collection of remains from the Viking Age and later periods. "It was literally thousands of fragments," she says. Yet they were just sitting there, hardly examined by anyone.

Hayeur Smith grew up surrounded by fabrics her anthropologist mother collected from around the world. In her 20s Hayeur Smith earned a fashion degree in Paris. She knew that the way people in the past clothed themselves and wove everything from currency to cloaks could reveal a great deal about a lost culture, especially its women. In the 1990s, as a Ph.D. student at the University of Glasgow, she'd devoted herself to studying Viking women's dress and ornament, typically from artifacts found in burial sites. Inspired by her first glimpse of the wealth of textile remnants in the museum's storeroom, Hayeur Smith eventually decided to uncover the lives of the ordinary women who stood weaving at their looms.

Ever since then, she has been analyzing textiles spanning 900 years of history, starting with the Viking settlement of Iceland in C.E. 874. She has pored over thousands of soil-encrusted fragments dense with information about the women who made the fabric. Her resulting studies of that museum's neglected collection of little brown scraps, as well as many other specimens of ancient Viking and later North Atlantic fabric, are among the first to prove the old guard wrong about the importance of cloth and women in ancient societies.

Textiles *trivial?* In my Zoom interview with her, Hayeur Smith, blond hair spilling to her waist, calling

© 2023 Scientific American



LEGAL CLOTH known as *vaðmál* was recovered from late 17th–century archaeological deposits at the site of Gilsbakki in western Iceland. Gilsbakki was the seat of Viking Age and medieval chieftains from around C.E. 900 to C.E. 1210.

to mind a Valkyrie, speaks in a voice ringing with conviction: "No. Textiles and what women made were as critical as hunting, building houses and power struggles," she says. In the Viking and medieval eras, women were the basis of the North Atlantic economy, and their cloth allowed people to survive the climate of the North Atlantic.

HIDDEN FIGURES

IN POPULAR CULTURE, Viking women are seen through the eyes of the era. In the 1950s they were portrayed as weak and subservient to men. In the 1970s they were sexualized. In recent shows such as *Vikings* and *The Last Kingdom,* they are depicted as shield-maidens or warriors.

Until Hayeur Smith began her work, the real lives of Viking women were largely unknown to science. According to archaeologist Douglas Bolender of the University of Massachusetts Boston, who studies the Viking Age and the medieval North Atlantic, the basic outline of Viking society came from the Icelandic sagas. Those book-length narrative accounts were set down more than 300 years after the events they describe. And the authors, who were men as far as we know, were Christianized people writing about their "pagan" ancestors.

Viking women have long been stereotyped in archaeology as performing primarily domestic tasks: child-rearing, cooking, weaving and making clothing. Written accounts and archaeological evidence confirm that they were weavers. Yet for years at a time during their husbands' absences for raids or trading expeditions, women ran the farms and engaged in trade, Hayeur Smith says.

"There's some truth" to the idea that we've found women's work less interesting, says archaeologist Thomas McGovern of the City University of New York. McGovern, whose full white beard evokes an Old Testament patriarch, entered archaeology in the 1970s. "Mostly it was old white guys," he recalls. Since then, however, the field has changed for the better, he says, with far more women and diversity generally.

Yet traditional views of women still color researchers' interpretations of evidence, says archaeologist Marianne Moen of the Museum of Cultural History in Oslo. A Viking expert who studies gender in the archaeological record, she says that she regularly sees how the meaning of artifacts is distorted by preconceptions of what they must signify. For example, a grave filled with a warrior's weapons at the Viking site of Birka in Sweden was long thought to be a man's final resting place until DNA evidence proved it was a woman's.

Alexandra Sanmark of the University of the Highlands and Islands in Perth, Scotland, an authority on Vikings and medieval archaeology, agrees. A man bur-

© Sarah Philbrick Kelly, 2008

© 2023 Scientific American

ied with scales is seen as a merchant, she says, but a woman buried with scales must be a merchant's wife, despite ample evidence that women conducted trade.

WRITTEN IN CLOTH

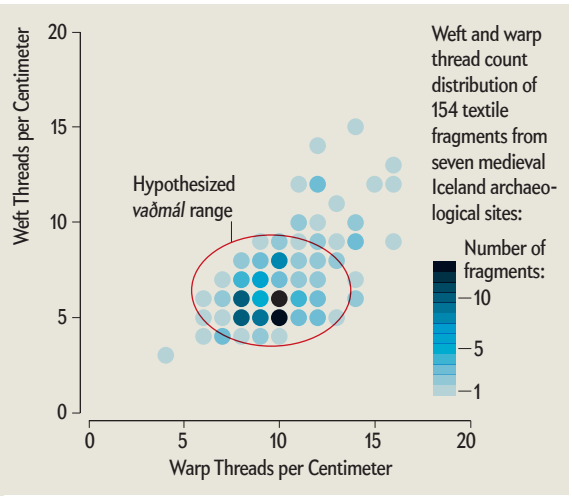
HAYEUR SMITH DECIDED to seek out North Atlantic women in the work of their hands. So little has been known about them until now, she says, "because it was men analyzing this from the perspective of men and medieval law codes written by men. Nobody had gone and looked at the actual stuff made by women."

She did not begin her textile analysis completely from scratch. There had been a few studies of textiles, most notably by the late Elsa Guðjónsson, whose work was foundational for Hayeur Smith. Guðjónsson, however, had been able to study only "a handful" of archaeological pieces of fabric from the mountain of artifacts in the Iceland museum's collection, Hayeur Smith says. And Guðjónsson's work, like that of other textile analysts, focused mainly on technical details such as thread counts, weave types, fleece varieties, embroidery stitches and tools used to make them to understand the techniques of weaving.

For Hayeur Smith, the technical details were important, but she had a different goal: to create what she calls a "social archaeology" of the culture through which she could uncover the lives of the women who created the cloth. For this purpose, she focused on the everyday "homespun"—plain woolen fabric—made by ordinary women, who left no elaborate graves on their farms throughout the North Atlantic. Their only memorials are the textiles they wove on their warp-weighted looms.

Hayeur Smith demonstrated the Vikings' style of weaving at an event organized by the Haffenreffer Museum of Anthropology at Brown in 2020, a recording of which is available on YouTube. A wood horizontal bar resting on two vertical ones holds the separate vertical warp threads, which are weighed down taut by volcanic stones of the kind that dot the shores of Iceland. Holding a heddle rod to separate the warp threads, she draws the continuous horizontal weft thread in and out of one or more warp threads. By varying the number of warp threads, weavers could create common Icelandic patterns, mostly basic weaves known as twills and tabbies.

Before about C.E. 1000, the loom would have been set up in a *dyngja,* a weaving hut, says University at Albany archaeologist Kevin Smith, Hayeur Smith's husband, who has excavated similar structures in Iceland. These pit houses, he explains, are dug down 1.5 to three feet deep, sometimes with turf walls above the pit and sometimes with wood walls that would have provided a space high enough for people to stand and work. With a stone-framed hearth in one corner to offer warmth and light, these small buildings—no more than nine by 15 feet in size—would have provided an intimate space fitting a loom and perhaps three women, spinning, weaving and sharing stories.

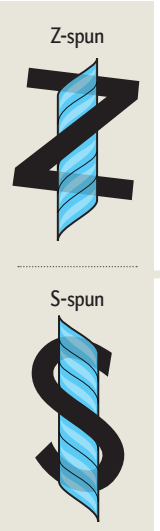


Hayeur Smith made several trips to the museum's basement laboratory in 2010, examining specimen after specimen under a microscope, counting warp and weft threads, taking note of such characteristics as the spin direction of yarn, the kind of fiber used and the weave type. As she worked, Hayeur Smith entered her data and extracted small samples for further analysis and testing, including a type of radiocarbon dating called accelerator mass spectrometry.

For the next several years she studied fabric remains from museum repositories in Iceland, Greenland, the Faroes, Scotland and Norway. Wielding her digital caliper, she measured the diameter of the cloth fibers and the sizes of the remnants. For each specimen, she meticulously recorded the age, site of origin and manufacturing details such as warp threads per square centimeter.

Somewhere between the first and second year of this endless and "filthy" job, soil all over her fingers, Hayeur Smith had her eureka moment. "Look," she shows me on a video call, holding her book open to a graph and pointing to a thick cluster of circled icons. "The more sites I checked, the more I saw this pattern. Viking Age textiles were colorful and varied, but in medieval times, there is a complete shift into standardized cloth."

Regular modern cloth, Hayeur Smith explains, could vary from 75 to 300 warp threads, but in Iceland and only in Iceland, in the 12th through 17th centuries, every textile from every site fell into a tight range of four to 15 warp threads. In addition, the spin direction of the yarn—clockwise (Z-spun) versus counterclockwise (S-spun)—shifted almost completely from Z-spun warp and weft to S-spun in the weft in the 11th century. All these details are specifications for legal cloth, called *vaðmál.* "Women were making the money!" she says.



CLOTH CURRENCY

ICELAND'S BUDDING ECONOMIC SYSTEM was based on Norway's. Certain commodities—cloth, cows, but-

Source: *The Valkyries' Loom: The Archaeology of Cloth Production and Female Power in the North Atlantic,* by Michele Hayeur Smith. University Press of Florida, 2020 (*chart reference*)

Graphics by Jen Christiansen

© 2023 Scientific American

ter, grain—were legally assigned a value based on their equivalent value in silver. Toward the end of the Viking Age, however, homespun woolen cloth became much more important as a form of exchange in Iceland than in Norway. Scholars believe this shift may have resulted from such factors as a scarcity of silver after the Vikings stopped raiding, population growth, and the colony's burgeoning wool production. "Although its value was still measured, in theory, against silver, this cloth … came to be legally regulated as an exchange good in and of itself," Hayeur Smith notes in her 2020 book *The Valkyries' Loom: The Archaeology of Cloth Production and Female Power in the North Atlantic*. The name *vaðmál,* she explains, is a combination of the Old Norse words *vað* ("stuff" or "cloth") and *mál* ("measure"), meaning "cloth measured to a standard." It is frequently mentioned as a measure and medium of exchange in Icelandic legal texts, as well as sales accounts, church inventories and farm registers, from the 1100s into the 17th century.

Women made all of the *vaðmál*. In fact, they were churning out huge quantities of it as both a unit of currency and a commodity to be sold near and far. *Vaðmál* could be used to pay taxes and tithes, but it could also be traded or sold for making clothes and other necessities. It was especially in demand in England, which produced its own luxury fabrics but needed great quantities of Iceland's cheap durable homespun to clothe peasants, the urban poor and common soldiers.

It's a modern idea that work done at home is "domestic" and lesser because it doesn't produce money, Moen says. In the North Atlantic world, "home was where work was done." In fact, as Hayeur Smith points out, *vaðmál* was a major income-generating product.

Scholars knew about *vaðmál* in an "abstract" sort of way, Hayeur Smith says, because it was precisely defined in the medieval law books. But the legal texts never mention the women weaving it, she points out. And nobody checked the cloth remains to see whether they conformed to the specifications in the legal texts.

In tandem with her textile analysis, she examined the legal texts—most of which had, thankfully, already been translated from Old Norse into modern English. Through painstaking inspection, she confirmed that the cloth Icelandic women wove conformed exactly to these standards: a 2/2 twill (a tweed), Z/S-spun, woven with four to 15 warp threads per centimeter. The cloth was also supposed to measure two "ells" in width and six ells in length (a little more than one yard wide and three yards long in modern measurements). Based on the fragments she analyzed, the Icelandic cloth "can be assumed to be about this size or bigger." That unit of cloth was equal to a certain weight of silver. "Everyone assumed the economy was a male thing," Hayeur Smith says. In fact, it was not the men but the women making the decisions.

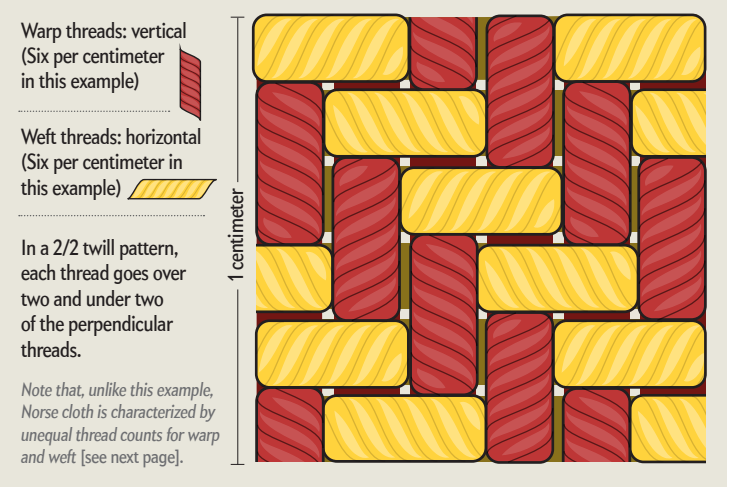
She suggests that women either created the specifications themselves or collaborated with men to do it. "It's hard to know what they were thinking," Hayeur Smith admits. "What's clear in looking at the time involved and the tools is that every able-bodied female in a household would have been involved." Indeed, at a certain level, "they may have controlled a lot of the narrative on these legal guidelines," she says. "It's not the men sitting there writing books … because [men] didn't go near weaving," she says.

LOOMING TABOOS

HAYEUR SMITH GROUNDS THIS assertion partly on evidence from poetic and mythological sources, including the Icelandic sagas, which provide clues to deep-seated attitudes toward women and weaving in the Viking Age and beyond. The power of women is expressed in the *Darraðarljóð* from Njáls saga, says Karen Bek-Pedersen, an expert on female aspects of Viking religion at Aarhus University in Denmark. In the saga, a soldier at the dawn of a battle has a vision in which he peeks into a *dyngja* and sees 12 Valkyries, Odin's female warrior spirits. They set up a loom and start weaving, using men's viscera as parts and threads. As they weave, they describe—and determine—the bloody defeat to come, Bek-Pedersen explains. She quotes this stanza:

The fabric is warped
with men's intestines
and firmly weighted
with men's heads;
bloodstained spears serve
as heddle rods,
the shed is ironclad,
and pegged with arrows.
With our swords we must strike
this fabric of victory

Poems from the sagas probably predate the sagas themselves, Bek-Pedersen says. Filled with metaphor,


Warp threads: vertical (Six per centimeter in this example)
Weft threads: horizontal (Six per centimeter in this example)
In a 2/2 twill pattern, each thread goes over two and under two of the perpendicular threads.
Note that, unlike this example, Norse cloth is characterized by unequal thread counts for warp and weft [see next page].
1 centimeter

© 2023 Scientific American

alliteration, rhythm and rhyme, they are hard to alter and easy to remember, making them likely to have been passed down through oral tradition.

The *dyngja,* Bek-Pedersen observes, can be seen as a space "charged with a feminine energy that reaches beyond the abilities of ordinary human women." In the literary canon, she says, men who hang out there and gossip with the women are portrayed as cowards or villains and invariably come to a bad end.

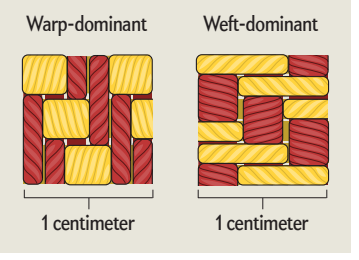
The fact that the *dyngja* was a space men shunned weighs heavily in Hayeur Smith's appraisal of women's power in cloth-making. Men feared that if they entered, they would lose their masculinity or even their lives. After Scandinavia turned Christian around C.E. 1000, looms were brought into the main living area of the *skáli,* the longhouse. Weaving would have been carried out in a separate area or room, with the taboos about this women's craft most likely undiluted. These taboos became a critical factor in women's power as their cloth turned into a major driver of the Icelandic economy.

CLOTH AND CLIMATE

IN 2011 HAYEUR SMITH met McGovern in a Chinese restaurant on the Upper East Side of Manhattan for what must have looked to an outside observer like a peculiar hand-off. McGovern had brought some fragile remains from an excavation he and his team had carried out a couple of years earlier at a site called Tatsipataa in southwestern Greenland. The workers had scooped up the textile bits separately from the bones he was studying, and he was happy to give them away, wondering what she could possibly learn from them. When he handed them over, he recalls, he thought, "Well, good luck with this!"

Hayeur Smith was on a quest to find out why the cloth made by Greenland's women diverged so much from the cloth made by Iceland's female weavers. Greenland had been settled in C.E. 986 entirely by Icelanders. They were followers of Erik the Red, who'd been exiled from Iceland for manslaughter. The Greenlanders' cloth started out identical to the Icelanders' warp-dominant fabric but eventually shifted to contain more threads in its weft than its warp.

In the early 2000s the late Else Østergård, a textile expert at the Danish National Museum, proposed an explanation for the shift. Holding up her lovingly tattered copy of Østergård's 2004 volume *Woven into the Earth,* Hayeur Smith says Else thought it possible that Greenlandic women's weaving innovations might be a response to climate change during a period of cold climate known as the Little Ice Age. In Greenland, the first dramatic drop in temperatures started around roughly 1340 and continued with fluctuations through the mid-15th century, when Greenland's colonies disappeared, to the 1900s.



Hayeur Smith set out to test Østergård's hypothesis against the archaeology, starting with the evidence from Tatsipataa. McGovern's specimens turned out to be "phenomenal," she says. Excavated under controlled conditions from a well-documented series of layers of remains, they were brimming with information about changes in weaving, when these changes occurred—and, quite possibly, *why.* Collaborating with Konrad Smiarowski, at the time one of McGovern's doctoral students, Hayeur Smith reviewed their excavation plan, a depiction of how the artifact layers were deposited over time. It indicated that the weft-dominant cloth appeared somewhat later in time.

By dating the Tatsipataa cloth remains, Hayeur Smith was able to correlate the ratio of weft to warp threads in each sample with published records of climate data. As Østergård had hypothesized, weft-dominant cloth did indeed increase as temperatures dipped in the 1300s. "It matched up perfectly with the climate data!" she says.

Still, as exciting as that finding was, "it was just one site," Hayeur Smith says. To prove that women were adapting their weaving to climate change, she had to collect remains from all over Greenland during different time periods.

In September 2014 Hayeur Smith's pursuit of women in textiles brought her to what had once been an old trading house on Greenland's Nuuk Harbor, where she could watch icebergs float by. Built by 18th-century Danish-Norwegian missionary Hans Egede, the building was now used by the Greenland National Museum and Archives to host research teams. But she was the only researcher there. At night the storms battering the house and the groaning beams evoked a sense of uneasiness that challenged her scientific outlook. Later she would learn from locals that this abode, situated beside a graveyard for smallpox victims, was widely regarded as haunted.

She laughs at her fears as she recounts them in 2022 from her cozy home office in Pawtucket, R.I., which is filled with artwork and portraits of her French Canadian and American great-grandfathers in antique oval frames. In the Nuuk museum and, during another trip that same year, in the National Museum of Denmark in Copenhagen, Hayeur Smith inspected some 700 cloth specimens from multiple archaeological sites across Greenland. She returned to Nuuk in 2017 to study even more samples. With all the dating she did, plus Østergård's, she was able to track the evolution of weft-dominant cloth and correlate it with the years of climate change. "I *confirmed* [it was] climate change," she says.

Flicking her long hair back from her face, Hayeur Smith points to a graph in her book. "Look, that's the climate data." She draws my attention to an arrow that goes down to the year 1320. "That's when you see

© 2023 Scientific American

weft-dominant cloth," she says. That type of cloth becomes widespread between 1300 and 1362.

After those dates, the weaving of weft-dominant textiles intensifies. Previously a minor type of cloth, "it becomes the most common textile produced in Greenland," she wrote in *The Valkyries' Loom*. It was "almost certainly" a response by local weavers in Greenland in this medieval era to cope with colder temperatures. Hayeur Smith had found her women. "I could see in the piece of cloth," she says, her voice buzzing with excitement, "the actual deliberate decision-making that women were doing, like *it's getting cold; let's change the way we weave our cloth*. It's almost unheard of that you get to see people's direct actions and thought processes" so far in the past.

POWER SHIFTS

EVENTUALLY, THOUGH, NATURAL, political and economic forces combined to strip Icelandic and Greenlandic women of the power they possessed as a result of making the all-important cloth. By around 1450 the Little Ice Age, among other factors, had destroyed the Greenland Norse colony, and plague and political upheaval had roiled the Kingdom of Norway.

In 1603 Danish authorities under King Christian IV imposed a royal monopoly on trading and strictly required all imports and exports to go through Denmark, reducing Iceland's freedom to trade. And although Iceland continued to use *vaðmál* as currency and export it until the late 17th century, fish had replaced the cloth as Iceland's primary export starting in the 14th century.

Imitating the English, whose male weaving guilds had produced fine cloths on foot-powered treadle looms since the 1300s, the Danes trained North Atlantic men to weave on these faster looms. They set up production workshops in locations around Iceland, including Reykjavik. They gave women spinning wheels, a much more efficient way to create yarn than the traditional spindle whorls used on the drop spindle. The Danes also encouraged women to knit—a skill they'd learned in the 1500s—responding to a market demand for knitted exports. In addition, they imported fabric from Denmark to Iceland. Women could buy it to make clothing, saving them the relentless labor of weaving. In taking these measures, the Danes essentially pushed women out of the mainstream of weaving.

Nevertheless, as Hayeur Smith found in the archaeological record, women continued to weave their homespun cloth on their farms. Fragments of the textile have surfaced at 17th- and 18th-century sites throughout Iceland, including its wealthiest ecclesiastical center, Skálholt. She believes people used it as a statement of national identity in the face of Danish rule and the new laws imposing a transformation of the women's 900-year-old tradition of textile production. "I see it as resistance," she says.

Yet the Danes—and the growth of industrialization—eventually prevailed. By the early 1800s, Hayeur Smith says, no one even knew how to weave on the old looms. And women were worse off for it. Once textiles could be made so much faster on machines than by hand, they came to be associated with things considered "frivolous or peripheral to our daily lives or of interest and important mainly to women as their primary consumers," she observes in *The Valkyries' Loom*. "It was in part the Industrial Revolution that sealed the fates of women as second-class citizens and ensured that Western society would become so vehemently patriarchal."

Hayeur Smith remains committed to the pursuit of the tales only cloth can tell. As the once dubious McGovern observes, "after her publications, nobody's going to look at textiles the same way again." SA

MICHÈLE HAYEUR SMITH examines archaeological deposits at Gilsbakki spanning 1,000 years—from the early 10th century to the early 20th century.

Francine Russo is a veteran journalist specializing in social sciences and relationships. She is author of *Love after 50: How to Find It, Enjoy It, and Keep It* (Simon and Schuster, 2021).

© Kevin Philbrook Smith, 2009

© 2023 Scientific American

METAL WORLD: Asteroid Psyche, shown in an artist's conception, may be the iron-nickel core of an early planet precursor.

Our neighborhood of planets was not created slowly, as scientists once thought, but in a speedy blur of high-energy crashes, destruction and rebuilding

By Linda T. Elkins-Tanton

© 2023 Scientific American

SOLAR SYSTEM SMASHUP

© 2023 Scientific American

I WAS WALKING OUT OF A CLASSROOM AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY, where students and I had been talking about the way planets form, when I was stopped by my colleague Ben Weiss. He studies magnetism in space rocks, and he was very excited. Weiss pulled me down the hall to his office to show me new data on one of these rocks, a meteorite called Allende. It was information that could change almost everything planetary geologists thought about the solar system. It was 2009, and that fall Weiss's research team had shown that Allende—which crashed into Earth in a huge fireball in Mexico in 1969 and contained some of the oldest known material in our system—harbored signs of an ancient magnetic field in its rock. The discovery was a surprise. This kind of field, astronomers thought, was made only by a magnetic dynamo of intensely hot, flowing liquid metal inside a planet, the way Earth's magnetic field is produced by liquid iron spinning in the planet's core. But Allende was supposed to be a fragment from a parent planetesimal—an early, nascent planet—that had been only slightly warm. Scientists presumed it had never gotten hot enough to melt the metal it contained. How, then, Weiss wondered, could this ancient piece of our solar system have become hot enough to create a magnetic dynamo?

My students had just been pelting me with questions about planetary evolution, challenging me to rethink some textbook wisdom, so I happened to have the bare bones of a new idea that might help answer Weiss's question. I went over to his whiteboard and began sketching it out.

Planetesimals had long been known to contain short-lived, unstable aluminum atoms radiating excess nuclear energy. This radioactive isotope is called 26Al, and when it decayed, that excess energy could have heated planetesimals. Conceivably, the heat from 26Al in Allende's parent body could have climbed so high that the object actually melted from the inside out. Metal within the body would have separated from silicate minerals in the rest of the rock and formed a liquid core that started to spin as the space rock rotated, creating a magnetic dynamo. Meanwhile the outside of the planetesimal would have been chilled by the cold of space, and cold rock and dust from our solar system's primitive disk would have kept adding to this unmelted rind.

This idea that early building blocks of the solar system contained so much energy was not the story I had learned in high school. Textbooks often still say that the solar system formed in a quiet, stately manner. That creation, 4.567 billion years ago, was thought to be an orderly, civilized process, like a minuet: the gas and dust of a molecular cloud spun down into a disk around a growing young star, and the gas and dust coalesced into many little boulders, each of which gradually grew to tens to hundreds of kilometers in diameter. These planetesimals themselves then rammed together to form larger bodies, each perhaps the size of Mars, called planetary embryos. Only after that did the temperature in this space nursery pick up. These embryos, which in growing had gained gravity strong enough to begin clearing their orbits of surrounding debris, then collided and grew into planets. Eventually components in these planets separated into the familiar churning metal core and a silicate mantle—hot, wildly volcanic places, antithetical to life.

That's the old view. By the time Weiss and I began pondering Allende, some other data, too, indicated that the earliest solar system was really a place of rapid and violent change. Now that tame sequence of dust to boulders to planetesimals to embryos to planets is being replaced. Actual planetesimal formation, once surmised to take place over hundreds of millions

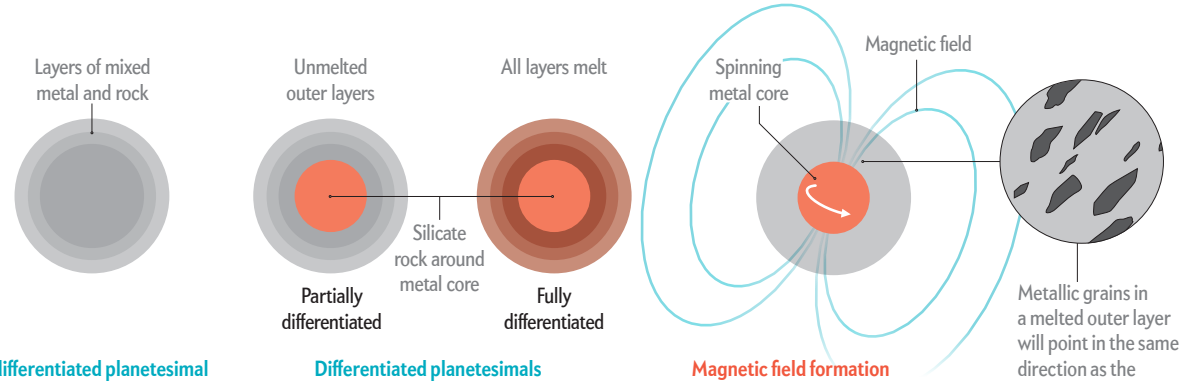
Courtesy of Asu and Peter Rubin/Iron Rooster Studios *(preceding pages)*

© 2023 Scientific American

Solar System Building Blocks

Planets like Earth and its neighbors got started when the cloud of gas and dust around our sun generated lumps known as planetesimals. These bodies were once thought to be quiet planetary seeds. But researchers now believe they crashed around and jump-started high-energy processes that gave rise to true planets.

Inside a Planetesimal


Layers of mixed metal and rock
Unmelted outer layers
All layers melt
Silicate rock around metal core
Partially differentiated
Fully differentiated
Spinning metal core
Magnetic field
Metallic grains in a melted outer layer will point in the same direction as the magnetic field and retain that same orientation after the dynamo has stopped and the outer layer has cooled.

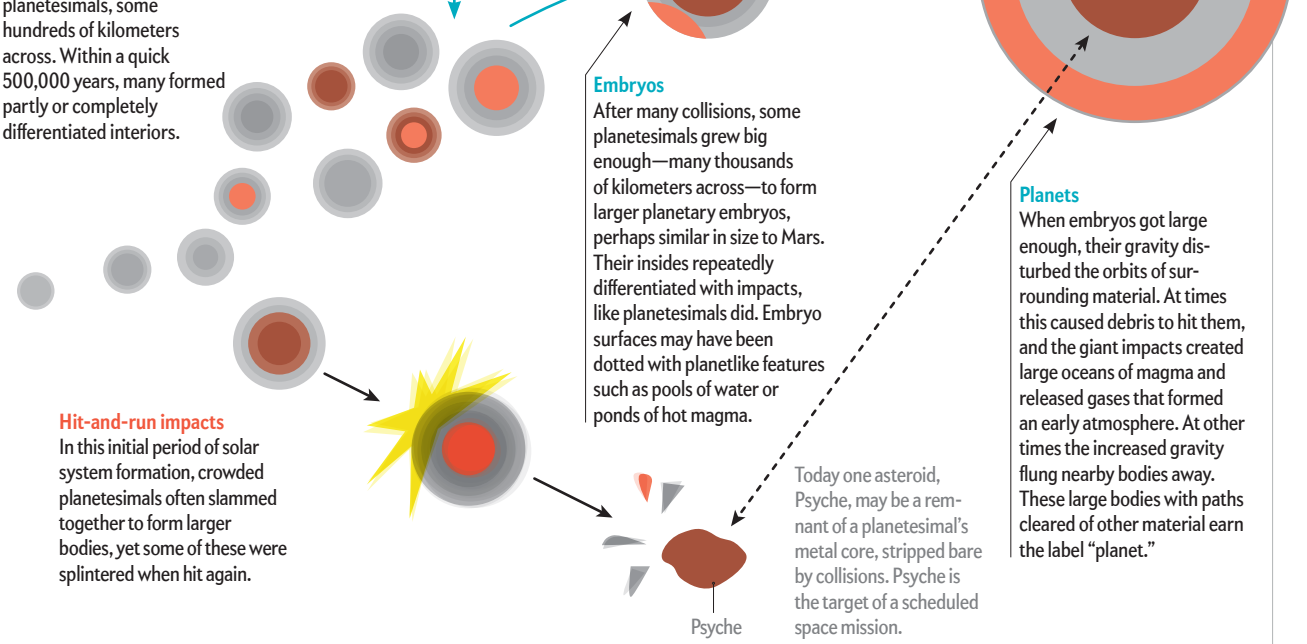
Undifferentiated planetesimal
These rocks were 10 to a few hundred kilometers in diameter and contained radioactive elements that heated their innards. They had an onionlike structure.

Differentiated planetesimals
Radioactive elements created more heat, and the body melted from the inside out. Denser iron-nickel metal sunk inward. Sometimes an outer rind was unaltered. All differentiated planetesimals had a liquid-metal core.

Magnetic field formation
Metal in motion can become a magnetic dynamo, which happened when planetesimals and their liquid metal cores started to spin. These cores generated magnetic fields pointing in particular directions, like Earth's field points north.

From Planetesimals to Planets

Accretion
When the solar system began, dust particles banged together, clumping into planetesimals, some hundreds of kilometers across. Within a quick 500,000 years, many formed partly or completely differentiated interiors.



Embryos
After many collisions, some planetesimals grew big enough—many thousands of kilometers across—to form larger planetary embryos, perhaps similar in size to Mars. Their insides repeatedly differentiated with impacts, like planetesimals did. Embryo surfaces may have been dotted with planetlike features such as pools of water or ponds of hot magma.

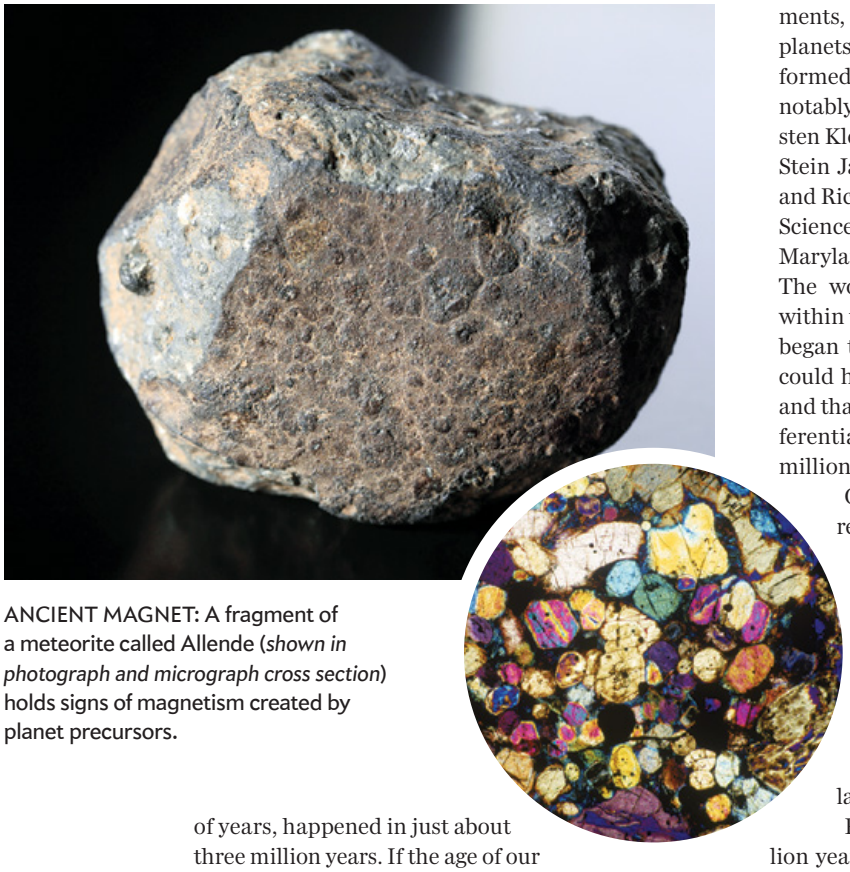
Hit-and-run impacts
In this initial period of solar system formation, crowded planetesimals often slammed together to form larger bodies, yet some of these were splintered when hit again.

Today one asteroid, Psyche, may be a remnant of a planetesimal's metal core, stripped bare by collisions. Psyche is the target of a scheduled space mission.

Planets
When embryos got large enough, their gravity disturbed the orbits of surrounding material. At times this caused debris to hit them, and the giant impacts created large oceans of magma and released gases that formed an early atmosphere. At other times the increased gravity flung nearby bodies away. These large bodies with paths cleared of other material earn the label "planet."

Illustration by Jen Christiansen

© 2023 Scientific American



ANCIENT MAGNET: A fragment of a meteorite called Allende (*shown in photograph and micrograph cross section*) holds signs of magnetism created by planet precursors.

of years, happened in just about three million years. If the age of our solar system, in human time, is now a day, this growing up happened during the very first minute. More energy in the early small components—energy like that provided by the aluminum heating and early collisions—means they did not have to wait and grow before developing different layers. Relatively minute planetesimals could host processes previously believed to be confined to planets, from melting and degassing to the creation of magnetic dynamos and volcanism.

And things in this system did not simply grow from small to large. Often large things blew apart into smaller pieces. Planet-size spheres may have formed in those early years through collisions of these smaller high-energy bodies, but glancing, hit-and-run blows among planetesimals sometimes stripped or destroyed them instead. Their debris could hit other spheres, enlarging them to planet size. Planets could be built, torn asunder and rebuilt in just 10 million years or less.

THE DISAPPEARING DISK

PLANETARY SCIENTISTS pieced together this novel, three-ring-circus vision of the active young solar system with great help from new tools for calculating the ages of meteorites and of planetary dust clouds—similar to our primordial solar system—elsewhere in the cosmos.

Over the past 15 to 20 years scientists have developed instruments capable of measuring the elements that make up space rocks to parts per million or even less. Because we have a good idea of how long it takes radioactive elements to decay into their daughter elements, such measurements allow us to date when the planets and planetesimals that shed these fragments formed and changed. Scientists all over the world—notably Alex Halliday of Columbia University; Thorsten Kleine of the University of Münster in Germany; Stein Jacobsen of Harvard University; Mary Horan and Rick Carlson, both at the Carnegie Institution for Science; and Richard Walker of the University of Maryland—have measured collections of meteorites. The work made clear that planetesimals formed within the first few million years after the dusty disk began to cool, that many of our terrestrial planets could have formed within the first 10 million years, and that even most of Earth probably formed and differentiated into a core and mantle by a few tens of millions of years.

Other avenues of research have yielded similar results. With ever improving telescopes, we can see young stars growing in other parts of the Milky Way, and in some cases, we can see the dust and gas disk from which a star and its planets grow. By estimating the ages of stars orbited by planets and comparing those measures with estimates for stars surrounded only by disks of dust and gas, researchers determined about 15 years ago that these disks last, on average, only three million years.

Planetesimals, therefore, have just three million years to grow on average. Any dust and gas not accreted onto rocks by then is lost, burned in the star or dispersed in space, and no more material for planet building is available. Considering that theorists used to think accretion took hundreds of millions of years, this is quite an acceleration!

More evidence for this timing comes from using the decay of radioactive elements like a clock that ticks away at a steady rate as one element turns into another. The new instruments gave teams throughout Europe and the U.S. enough precision to measure these elements and thereby learn how long that clock had been ticking. Meteorites that have fallen to Earth contain those elements. Most are pieces of asteroids, which themselves are primitive remnants of planetesimals. (Some other meteorites are from the moon, some are from Mars and others are from bodies not yet identified.)

One radioactive isotope of the element hafnium prefers to stay in silicate minerals, such as those in Earth's mantle. But it decays to an isotope of tungsten, which combines readily with metals that make up planet cores. This decay happens on a regular schedule: half the hafnium turns into tungsten in nine million years. This system gives us the timescale of metal-silicate (core-mantle) differentiation of planets and planetesimals: metal core formation scavenges tungsten from the silicate mantle and carries it into the growing core. Any hafnium stays happily in the mantle and keeps on decaying into tungsten, which then remains in the silicate mantle if no further core formation occurs. When scientists measure the ratios of hafnium to tungsten in

Detlev van Ravenswaay/Science Source (*photograph*); Michael Abbey/Science Source (*micrograph cross section*)

© 2023 Scientific American

meteorites, the amount of that tungsten isotope in them gives the time since core formation.

Such isotopic measurements of iron meteorites—many of which presumably hail from the metallic cores of planetesimals—show that their parent bodies formed within just 500,000 years of the first solids condensing from our protoplanetary disk of dust and gas. That time is less than 10 seconds into our putative 24-hour solar system. If iron meteorites are core fragments of planetesimals smashed apart by impacts, then planetesimals must have actually formed, melted and created iron cores within that tiny period.

GROWTH SPURT

WITH EXPERIMENTAL DATA firmly showing that the solar system formed much more quickly than textbook scenarios indicated, scientists who study it had to explain how planets came into being so quickly. That put the ball in the theorists' court. How can dust and pebbles just microns to centimeters in diameter, all orbiting around our young sun, clump together to form bodies up to 10 million times larger (100-kilometer planetesimals) in just 500,000 years?

With experimental data firmly showing that the solar system formed much more quickly than textbook scenarios indicated, scientists had to explain how planets came into being so quickly. That put the ball in the theorists' court.

The answer is not obvious. Standard physics says small clumps of dust that collide can readily stick together through electromagnetic forces, much the way static electricity creates clumps of household dust. The energy absorbed as a result of compression and loss of porosity during collisions also helps clumps stick together rather than bouncing off one another or breaking apart. As the clumps grow, though, they reach what is called the meter barrier. Before they reach a meter in diameter, these growing boulders become too large to stick by electromagnetic forces while still being too small to stick through gravitational attraction. Impacts even at very low speeds cause these conglomerates to destruct rather than accrete. Yet we know that such materials must be able to grow from meter size to planetesimal size—the planet we stand on today is a testament to that. So some other processes must have been at work.

Several ideas for how growth beyond the meter barrier occurs have been proposed. Most hypotheses involve concentrating material in the protoplanetary disk through various types of turbulence that slam particles together. Such whirlpool-like forces might include phenomena called Kelvin-Helmholtz eddies, which develop between the gas and dust layer of the disk and could effectively crush regions of material together into larger bodies. Much of that work has been pioneered by Anders Johansen, now at Lund University in Sweden. Hal Levison of the Southwest Research Institute and Johansen have separately worked on another model, called pebble accretion. His calculations indicate that even the smallest pieces of dust and clumps can be gravitationally diverted over several orbits to add to a growing planetesimal and can do so fast enough to build planetesimals early in the life of the solar system.

MELTING IN A FREEZER

NO KIND OF CRUSHING, however, could have caused the planetesimals to differentiate into mantles and cores. If planetesimals were first formed of primordial disk material in which metals and silicates were intimately mixed, then only high temperatures and at least partial internal melting would allow the metal to sink to the interior and form a core. Calculations show that slamming these relatively small bodies together would not provide enough energy to melt them. Consequently, researchers were left wondering where that melting energy—within the vast freezer of chilly space—came from.

This is where ideas about radioactive aluminum come in. Each time one of these atoms decays, it gives off a burst of heat. These tiny quanta of heat could have added up to a powerful energy source in the early solar system. Because aluminum is one of the six most common elements in stony material (the others are silicon, magnesium, iron, oxygen and calcium), 26Al, which has a half-life of about 700,000 years, could easily have heated at least some planetesimals to a melting temperature.

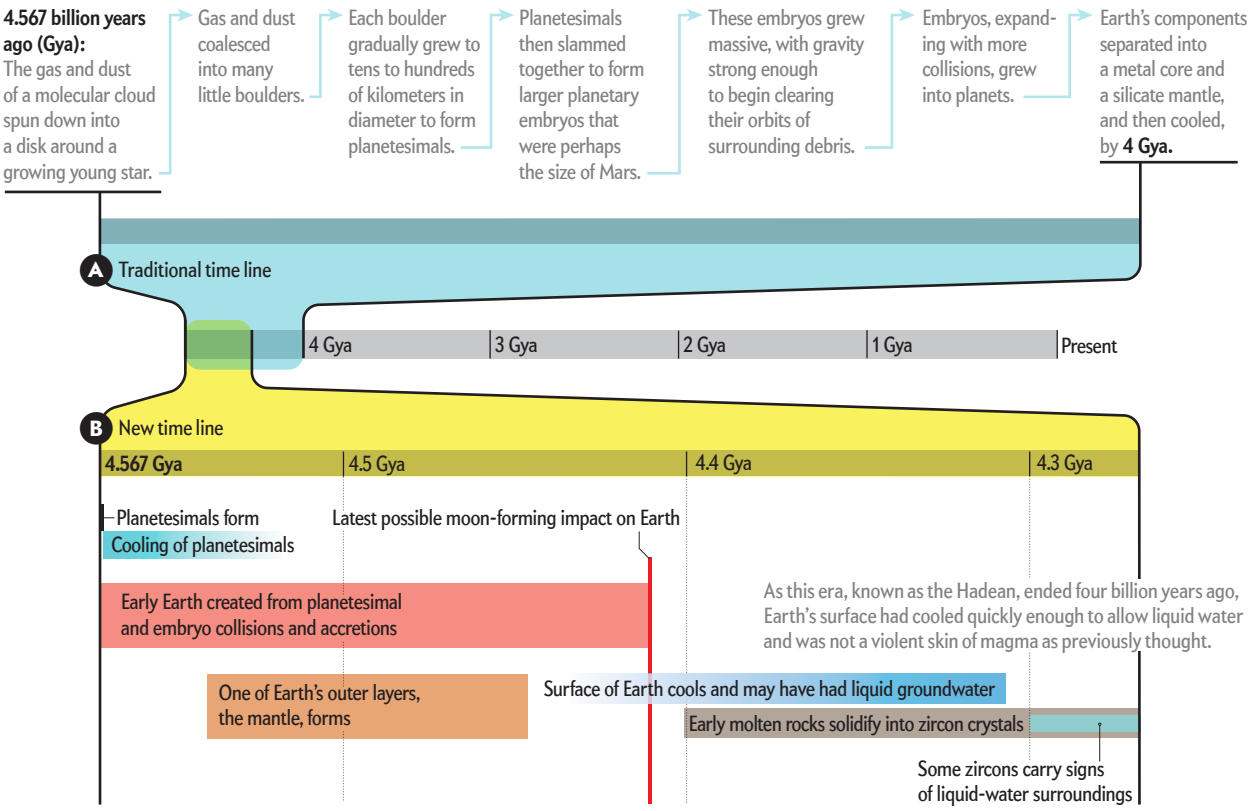
But what would have kept them from melting completely, considering that the recent observations strongly indicate that some of these planetary seeds had unmelted rinds on the outside? Size is part of the answer. In the case of Allende's parent planetesimal, to reach melting, the rocky body would have needed to be massive enough so that its interior produced heat faster than its exterior radiated it away. A larger body can get hotter inside than a smaller body can because it has a greater heat-producing volume compared with its heat-releasing surface. But the short half-life of 26Al means this growth had to be quick. To retain enough heat for the melting pattern we envisioned, Allende's parent planetesimal would have grown to 10 kilometers or more in radius within about two million years of the first solids in the solar system (the equivalent of the first 37 seconds of our 24-hour solar system)—and we think it may have grown to as much as 200 kilometers in radius.

It had been thought that planetesimals either melted entirely or remained primitive. But Weiss and I were suggesting a hybrid, where the most primitive material in the solar system encased a planetesimal that had

© 2023 Scientific American

Speeding Up Construction

Scientists used to think the formation of planets in our solar system was a slow, steady process that stretched over 500 million years, beginning at about 4.5 billion years ago **A**. Newer evidence from meteorites that fell to Earth, decay rates of radioactive elements and observations of dust around stars indicate our system—including a nascent Earth—got started much more quickly **B**, within three million years, in a high-energy blur of protoplanetary crashes and rebuilding.


4.567 billion years ago (Gya): The gas and dust of a molecular cloud spun down into a disk around a growing young star.
Gas and dust coalesced into many little boulders.
Each boulder gradually grew to tens to hundreds of kilometers in diameter to form planetesimals.
Planetesimals then slammed together to form larger planetary embryos that were perhaps the size of Mars.
These embryos grew massive, with gravity strong enough to begin clearing their orbits of surrounding debris.
Embryos, expanding with more collisions, grew into planets.
Earth's components separated into a metal core and a silicate mantle, and then cooled, by 4 Gya.
A Traditional time line
4 Gya
3 Gya
2 Gya
1 Gya
Present
B New time line
4.567 Gya
4.5 Gya
4.4 Gya
4.3 Gya
Planetesimals form
Cooling of planetesimals
Latest possible moon-forming impact on Earth
Early Earth created from planetesimal and embryo collisions and accretions
As this era, known as the Hadean, ended four billion years ago, Earth's surface had cooled quickly enough to allow liquid water and was not a violent skin of magma as previously thought.
One of Earth's outer layers, the mantle, forms
Surface of Earth cools and may have had liquid groundwater
Early molten rocks solidify into zircon crystals
Some zircons carry signs of liquid-water surroundings

melted inside—an outer rind as well as a melted core. This made sense because the Allende meteorite—with its record of a magnetic field caused by a heated interior—consists of unheated, primitive material alone. The only place it could come from would be a cool outer rind. Allende's parent planetesimal retained this primitive unmelted surface rind because it was chilled by the coldness of space and because dust in the cool protoplanetary disk continued to attach to it over time. Unchanged by heating, the rind was able to maintain a record of the magnetic field produced by the inner part of the planetesimal's structure, the melted core and its magnetic dynamo.

We were not the first to have thought of partial differentiation. Geologist John Wood drew a similar structure by hand in his Ph.D. thesis at M.I.T. in 1958—but no one had ever been heretical enough to say that the poster child for primitive unmelted meteorites, Allende, could have come into existence in this way or that the process was common, indeed formative, at the beginning of our solar system.

Now it does appear common. Researchers have found that at least four other meteorite parent bodies hosted magnetic core dynamos. At the same time, other possible sources of the magnetization have been ruled out: Allende and its fellows were not magnetized by a field produced in the sun, or by the dusty disk itself, or by transient plumes around impacts. If the earliest solar system was truly populated by hundreds or even thousands of differentiated planetesimals, zipping around and generating intense heat and magnetic dynamos like tiny Earths, it implied the entire infant system contained a great deal more heat than geologists had ever thought.

A CROWDED FIELD

OTHER IDEAS have also weakened the traditional notion of linear planet growth from small to large. For years, for the sake of simplicity and numerical tracta-

Graphic by Jen Christiansen

© 2023 Scientific American

bility, all simulations of planet formation assumed that in every collision of planetesimals, all the material from both colliders combined to form a new, bigger body. This merging occurred even as planetesimals were forming from dust. But more recent understanding and approaches to modeling the collision process were pioneered by Erik Asphaug of the University of Arizona. Some collisions are constructive, Asphaug found, and produce larger bodies. But other crashes can be destructive, with the collider stripping material from the target body and continuing on its way to do more damage elsewhere.

Only at about 10 million years of age did bodies really get bigger and stay big. What gave them enough stability to survive? Again, the answer seems to be size. As planetesimals collided and formed larger planetary embryos, their mass and therefore their gravity became larger. The gravity was great enough that whenever their orbits brought them near another object, that object was either drawn in and accreted through the pull of gravity or flung away as its orbit was changed. Thus, these growing planets began to clear out their orbits, which is one of the criteria for being properly called a planet. Smaller bodies had fewer and fewer refuges where their orbits could remain stable and unperturbed by growing planets; the asteroid belt became one of the safe havens left to them.

We wanted to explore a place that could prove or overturn hypotheses. The best destination, we had decided, was a world made out of metal: the metal asteroid Psyche. There is nothing quite like this place—at least nothing close enough to reach in a reasonable amount of time.

MISSION TO A METAL WORLD

WEISS, ASPHAUG, I and others would like to learn how our own planet's structure and composition came into being out of this energetic, often chaotic environment. But we cannot, despite the science fiction of popular movies or Jules Verne novels, get a good look at Earth's core. It is too deep underground and has pressures too high to sample directly.

But perhaps one particular asteroid, a remnant of an ancient planetesimal, might be a decent stand-in. About 12 years ago several of my colleagues and I began to design a space mission to explore this possibility. We gathered in the Left Field mission-formulation room at NASA's Jet Propulsion Laboratory. This room is designed for creativity. It has shelves full of drawing and building materials such as cardboard, wheels, wire, Legos, paper, markers and foam. It is a good room to think about something completely new, and that is what we were trying to do. We wanted to explore a place that could prove or overturn hypotheses. The best destination, we had decided, was a world made out of metal: the metal asteroid Psyche.

There is nothing quite like this place—at least nothing close enough to reach in a reasonable amount of time. Psyche is one of the largest asteroids, about 200 kilometers in diameter, and is located between Mars and Jupiter. All the physical measurements we have—from radio telescopes bouncing waves off the body—indicate that it is composed of significant amounts of iron and nickel. Psyche looks like it is a stripped-naked planetesimal core, a last remnant of the hit-and-run collisions that disrupted bodies in the early solar system. The orientation of particles in Psyche, like tiny magnetic compass needles, might tell us whether it had a magnetic dynamo. There might also be some remnants of its rocky exterior that tell us what the deep mantle of planetesimals looked like. If there were surface impacts on the naked metal, the splashes might have produced sharp metal cliffs that froze before they could fall back to the surface.

Each of us in the mission room brought a particular set of skills to the table: Weiss, his specialty of measuring magnetic fields in meteorites; William F. Bottke, the dynamics of orbiting bodies; Asphaug, the effects of collisions; Bruce Bills, calculating the gravity field of a body; Daniel Wenkert, managing data and operations. Damon Landau calculated trajectories; he is an interplanetary travel agent. John Brophy organized our deliberations, and I brought my knowledge of compositions, melting, solidification and differentiation processes.

The energy in the room rose to a heady peak. There was no e-mail checking, and there were no conversational pauses. We were united in our excitement over true exploration: humankind has never visited a metal body, and we do not even know what it will look like.

Since then, those days of playing with foam and Legos have given way to an organized push by a team of about 300 people, with more than 1,500 participating in the mission over the years. We proposed a spacecraft about the size of a singles tennis court, powered by solar cells and ion thrusters and carrying a magnetometer to detect magnetic fields, a gamma-ray spectrometer to identify elements and two cameras. In January 2017 we received the good news that NASA had selected the Psyche mission for flight. We will launch the spacecraft in October 2023, and we project that it will arrive at this strange remnant of planet building in 2029. The world of metal awaits. SA

Linda T. Elkins-Tanton, a planetary geologist specializing in the evolution of terrestrial planets, is vice president of Arizona State University's Interplanetary Initiative and principal investigator of the NASA Psyche mission.

© 2023 Scientific American

BREAKING COSMOLOGY

JWST's first images include unimaginably distant galaxies that challenge theories of how quickly these structures can form

By Jonathan O'Callaghan

© 2023 Scientific American

GALAXIES from the depths of cosmic time appear in a small crop from "deep field" observations taken by the James Webb Space Telescope (JWST). The most distant objects in such images may reveal surprising new details about the early universe.

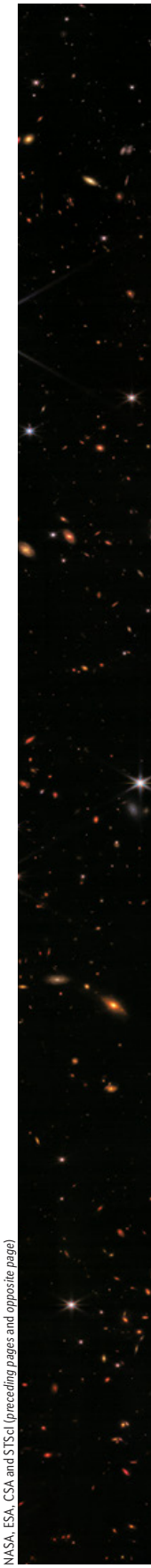
© 2023 Scientific American

ROHAN NAIDU WAS AT HOME WITH HIS GIRLFRIEND WHEN HE found the galaxy that nearly broke cosmology. As his algorithm dug through early images from the James Webb Space Telescope (JWST) late one night in July 2022, Naidu shot to attention. It had sifted out an object that Naidu recognized was inexplicably massive and dated back to just 300 million years after the big bang, making it older than any galaxy ever seen before. "I called my girlfriend over right away," Naidu says. "I told her, 'This might be the most distant starlight we've ever seen.'" After exchanging excited messages with one of his collaborators "with lots of exclamation marks," Naidu got to work. Days later they published a paper on the candidate galaxy, which they dubbed "GLASS-z13." The Internet exploded. "It reverberated around the world," Naidu says.

The discovery of this galaxy, just weeks into JWST's full operations, was beyond astronomers' wildest dreams. JWST—the largest, most powerful observatory ever launched from Earth—was built to revolutionize our understanding of the universe. Stationed 1.5 million kilometers away from earthly interference and chilled close to absolute zero by its tennis court–sized sunshade, the telescope carries a giant segmented mirror and exquisitely sensitive instruments that were designed to uncover details of cosmic dawn never before observed.

This is the scarcely probed era—no more than a few hundred million years after the big bang—in which the very first stars and galaxies coalesced. How exactly this process unfolded depends on exotic physics ranging from the uncertain influences of dark matter and dark energy to the poorly understood feedbacks among starlight, gas and dust. By glimpsing galaxies from cosmic dawn with JWST, cosmologists can test their knowledge of all these underlying phenomena—either confirming the validity of their best consensus models or revealing gaps in understanding that could herald profound new discoveries.

Such observations were supposed to take time; initial projections estimated the first galaxies would be so small and faint that JWST would find at best a few intriguingly remote candidates in its pilot investigations. Things didn't quite go as planned. Instead, as soon as the telescope's scientists released its very first images of the distant universe, astronomers such as Naidu (at the Massachusetts Institute of Technology) started finding numerous galaxies within them that, in apparent age, size and luminosity, surpassed all predictions. The competition for discovery was fierce: with each new day, it seemed, claims of yet



NASA, ESA, CSA and STScI (*preceding pages and opposite page*)

© 2023 Scientific American

THE INTERACTING galaxies of Stephan's Quintet, as seen by JWST, approximately 290 million light-years away from Earth. Covering one fifth of the moon's diameter, this mosaic is constructed from almost 1,000 separate images and reveals never-before-seen details of this galaxy group.

© 2023 Scientific American

NASA, ESA, CSA and STScI

another record-breaking "earliest-known galaxy" emerged from one research group or another. "Everyone was freaking out," says Charlotte Mason, an astrophysicist at the University of Copenhagen. "We really weren't expecting this."

In the weeks and months following JWST's findings of surprisingly mature "early" galaxies, theorists and observers scrambled to explain them. Could the bevy of anomalously big and bright early galaxies be illusory, perhaps because of flaws in analyses of the telescope's initial observations? If genuine, could they somehow be explained by standard cosmological models? Or, just maybe, were they the first hints that the universe is more strange and complex than even our boldest theories had supposed?

At stake is nothing less than our very understanding of how the orderly universe we know emerged from primordial chaos. JWST's early revelations could rewrite the opening chapters of cosmic history, which concern not only distant epochs and faraway galaxies but also our own existence here in the familiar Milky Way. "You build these machines not to confirm the paradigm but to break it," says JWST scientist Mark McCaughrean, a

© 2023 Scientific American

AN IMAGE FROM JWST reveals hundreds of previously invisible newborn stars in the stellar nursery known as the Carina Nebula, a vast agglomeration of gas and dust some 7,600 light-years from Earth.

senior adviser for science and exploration at the European Space Agency. "You just don't know how it will break."

DEEP LOOKS FOR COSMIC DAWN

ONE MIGHT SAY JWST'S observations of early galaxies have been billions of years in the making, but more modestly they trace back to the Space Telescope Science Institute (STScI) in 1985. At the time the Hubble Space Telescope was still five years away from launching on a space shuttle. But Garth Illingworth, then deputy director of the STScI, was surprised one day when his boss, then director Riccardo Giacconi, who died in 2018, asked him to start thinking about what would come after Hubble much farther down the road. "I protested, saying we've got more than enough to do on Hubble," Illingworth recalls. But Giacconi was insistent: "Trust me, it'll take a long time," he

© 2023 Scientific American

said. So Illingworth and a handful of others got to work, drawing up concept ideas for what became known as the Next Generation Space Telescope (NGST), later renamed JWST after a former NASA administrator.

Hubble would be transformational, but astronomers knew its capabilities would be limited by its observations in visible light. As light from a very distant galaxy travels across the cosmic abyss, it is stretched by the expansion of the universe—a broadening of wavelengths known as redshift. The higher the redshift value, the more stretching the light has experienced, and thus the more distant its source galaxy. Redshifts for early galaxies are so high that their emitted visible light has stretched into infrared by the time it arrives at our telescopes, which is why Hubble could not see them. The NGST, in contrast, would observe in infrared and would boast a very large (and very cold) starlight-gathering mirror, allowing it to peer much deeper into the universe. "Everybody realized that Webb would be the telescope for looking at early galaxies," Illingworth says. "That became the primary science goal."

The need for the telescope was highlighted in December 1995, when astronomers pointed Hubble at a seemingly empty patch of the sky for 10 consecutive days. Many experts predicted the extended observation would be a waste of resources, revealing at best a handful of dim galaxies, but instead the effort was richly rewarded. The resulting image, the Hubble Deep Field, showed the "empty" spot was filled with galaxies by the thousands, stretching back 12 billion years into the 13.8-billion-year history of our universe. "There were galaxies everywhere," says Illingworth, now an astrophysicist at the University of California, Santa Cruz. The Hubble Deep Field showed that the early universe was even more crowded and exciting than most anyone had expected, offering observational treasures to those who took the time and care to properly look. Yet, impressive as Hubble's Deep Field was, astronomers wanted more.

After more than two decades of labor at a cost of some $10 billion, JWST finally launched on Christmas Day 2021. The telescope reached its deep-space destination a month later, where it would endure exhaustive testing to ensure its optimal performance. By July 2022 it was ready to begin its long-awaited first year of science observations, known as Cycle 1. Part of the telescope's early time was devoted to high-impact programs across a range of disciplines from which data would immediately be made public. Two of those, CEERS (the Cosmic Evolution Early Release Science Survey) and GLASS (the Grism Lens–Amplified Survey from Space), independently spent dozens of hours looking for galaxies in the early universe by staring at separate small portions of the sky. Not much was expected—perhaps a slightly more ornate version of the Hubble Deep Field but nothing more. Steven Finkelstein of the University of Texas at Austin, the lead on CEERS, says extremely distant galaxies were predicted to pop up only "after a few cycles of data" from multiple programs.

Instead, much to the surprise of astronomers, extremely distant galaxies came into view immediately. Hubble's record for the most distant known galaxy had been GN-z11, spotted in 2015 at a redshift of 11 thanks to a 2009 upgrade that enhanced the telescope's modest infrared capabilities. A redshift of 11 corresponds to a cosmic age of about 400 million years, a point at the brink of when galaxy formation was thought to begin. But

NASA, ESA, CSA and STScI

from the very first GLASS data, two teams—one led by Naidu in that breathless late-night discovery—independently found GLASS-z13 at a redshift of 13, some 70 million years farther back in time.

In their quest for quick results, the researchers relied on redshift estimates derived from simple brightness-based measurements. These are easier to obtain but less precise than direct measurements of redshift, which require more dedicated observation time. Nevertheless, the simplified technique can be accurate, and here it suggested a galaxy that was unexpectedly bright and big, already bearing a mass of stars equivalent to a billion suns, just a few hundred times less than that of the Milky Way's stellar population, despite our own galaxy being billions of years more mature. "This was beyond our most optimistic expectations," says Tommaso Treu, an astronomer at the University of California, Los Angeles, and the lead on GLASS.

The record didn't last long. In the following days, dozens of galaxy candidates from CEERS and GLASS sprang into view with estimated redshifts as high as 20—just 180 million years after the big bang—some with disklike structures that were not expected to manifest so early in cosmic history. Another team, meanwhile, found evidence for galaxies the size of our Milky Way at a redshift of 10, less than 500 million years after the big bang.

Such behemoths emerging so rapidly defies expectations set by cosmologists' standard model of the universe's evolution. Called Lambda CDM (LCDM), this model incorporates scientists' best estimates for the properties of dark energy and dark matter, which collectively act to dominate the emergence of large-scale cosmic structures. ("Lambda" refers to dark energy, and "CDM" refers to dark matter that is relatively sluggish, or "cold.") "Even if you took everything that was available to form

© 2023 Scientific American

A SIDE-BY-SIDE COMPARISON shows JWST's remarkably detailed observations of the Southern Ring Nebula in near-infrared light (*left*) and mid-infrared light (*right*). Located more than 2,000 light-years from Earth, the nebula is composed of shells of gas and dust expelled from a dying star, which in each image can be seen near the nebula's core.

stars and snapped your fingers instantaneously, you still wouldn't be able to get that big that early," says Michael Boylan-Kolchin, a cosmologist at the University of Texas at Austin. "It would be a real revolution."

HOW TO BUILD A GALAXY

TO UNDERSTAND THE DILEMMA, a brief refresher is needed. In the first second after the big bang, our universe was an almost inconceivably hot and dense soup of primordial particles. Over the next three minutes, as the cosmos expanded and cooled, the nuclei of helium and other very light elements began to form. Fast-forward 400,000 years, and the universe was cold enough for the first atoms to appear. When the universe was about 100 million years old, theorists say, conditions were finally right for the emergence of the first stars. These giant fireballs of mostly hydrogen and helium were uncontaminated by heavier elements found in modern-day stars, so they possessed significantly different properties. Larger and brighter than today's stars, these first suns coalesced in protogalaxies—clusters of gas that clung to vast, invisible scaffolds of dark matter. Gravity guided the subsequent interactions between these protogalaxies, which eventually merged to form larger galaxies. This process of becoming, of the early universe's chaos giving way to the more orderly cosmos we know today, is thought to have taken about a billion years.

JWST's discovery of bright galaxies in the early cosmos challenges this model. "We should see lots of these little protogalactic fragments that have not yet merged to make a big galaxy," says Stacy McGaugh, a cosmologist at Case Western Reserve University. "Instead we're seeing a few things that are already big galaxies." Some of these galaxies may be impostors, much closer galaxies shrouded in dust that makes them look dimmer and farther away when brightness-based measurements are used. Follow-up observations of GLASS-z13 in August 2022 by the Atacama Large Millimeter Array (ALMA) in Chile, however, suggested that is not the case for this candidate, because ALMA did not see evidence for large amounts of dust. "I think we can exclude low-redshift interlopers," says Tom Bakx, an astronomer at Nagoya University in Japan, who led the observations. Yet the lack of dust means ALMA struggled to see the galaxy at all, showing how difficult it could be for telescopes to confirm observations made using JWST's advanced capabilities. "The good news is there's nothing detected," Naidu says. "The bad news is there's nothing detected." Only JWST, in this case, can follow up itself.

The most startling potential explanation is that the canonical LCDM cosmological model is wrong and requires revision. "These results are very surprising and hard to get in our standard model of cosmology," Boylan-Kolchin says. "And it's probably not a small change. We'd have to go back to the drawing board." One controversial idea is modified Newtonian dynamics (MOND), which posits that dark matter does not exist and that effects attributed to it can instead be explained by large-scale fluctuations in gravity. JWST's observations could support such a theory. "MOND has had a lot of its predictions come true; this is another one of them," says McGaugh, who is one of the idea's leading proponents. Others remain unconvinced. "So far everything that we've tried to test MOND hasn't been able to really provide a satisfactory answer," says Jeyhan Kartaltepe, an astrophysicist at the Rochester Institute of Technology.

One simpler possible solution is that galaxies in the early universe might have little or no dust, making them appear brighter. This scenario could confound efforts to calculate the galaxies' true masses and could perhaps also explain ALMA's difficulty spotting GLASS-z13. "It could be that supernovae didn't have enough time to produce the dust, or maybe in the initial phases [of galaxy formation] the dust is expelled from galaxies," says Andrea Ferrara, an astronomer at the Scuola Normale Superiore in Italy, who has proposed such a possibility. Mason and her colleagues suggest that, alternatively, in its observations of the early universe JWST may so far be seeing only the very brightest young galaxies, which should be the easiest to spot. "Maybe there's something happening in the early universe that means it's easier for some galaxies to form stars," she says.

David Spergel, a theoretical astrophysicist and current president of the Simons Foundation in New York City, agrees. "I think what we're seeing is that high-mass star formation is very efficient in the early universe," he says. "The gas pressures are higher. The temperatures are higher. That has an enormous impact on the environment for star formation." Magnetic fields might have arisen earlier in the universe than we thought, driving material to kick-start the birth of stars. "We might be seeing a signature of magnetic fields emerging very early in the universe's history," Spergel says.

© 2023 Scientific American

A RUSH TO BREAK THE UNIVERSE

THE RAPID FLOW of scientific papers from JWST's initial observations is no fluke; when the first data started streaming down, astronomers were eagerly waiting. "People had been working on their pipelines for years," Boylan-Kolchin says. Instead of the traditional peer-review processes, which can take months, astronomers published on arXiv, a website where scientific papers can be uploaded after minimal review by moderators but well before formal peer review. A new form of peer review unfolded in near real time on Twitter and other social media platforms. "It's science by arXiv," Naidu says. The resulting frenzy was intense—and surprising. "I expected a lot of activity," says Nancy Levenson, STScI's interim director. "But I underestimated the amount."

The result was that findings could be quickly publicized and discussed, but some fear at a cost. "People were rushing things a little bit," says Klaus Pontoppidan, JWST's project scientist at STScI. "The gold standard is a refereed, peer-reviewed paper." Early calibration issues with JWST, for example, may have affected some results. Nathan Adams of the University of Manchester in England and his colleagues found there could be dramatic changes, with one galaxy identified at a redshift of 20.4 recalibrated to just 0.7. "We need to calm down a little bit," Adams says. "It's a bit too early to say we've completely broken the universe."

Such issues are unlikely to eradicate all of JWST's high-redshift galaxies, however, given their sheer number. "It's more likely that the early universe is different from what we predicted," Finkelstein says. "The odds are small that we're all wrong." Astronomers are now racing to conduct follow-up observations with JWST. Levenson says she's currently reviewing about a dozen proposals from various groups asking for additional JWST observing time, most of which are seeking to scrutinize high-redshift galaxy candidates. "Considering the excitement and importance of these early discoveries, we thought it was appropriate to ask for a little bit of time to confirm them," says Treu, who put forward one of the proposals.

Other programs are now set to hunt for distant galaxies, such as COSMOS-Web, led by Kartaltepe, which is expected to hugely increase the known population of early galaxies by observing a wider swath of sky for hundreds of hours. "We estimate there are thousands we'll be able to detect," she says. Future projects might look for evidence of those first protogalaxies, perhaps using the explosive deaths of supersized first stars in especially luminous and energetic supernovae as markers for their existence. Some estimates suggest JWST could see as far as a redshift of 26, just 120 million years after the big bang, a cosmic blink of an eye. Much other work will be done to follow up the growing list of high-redshift candidates. "Even confirming a handful of these would be quite amazing," Naidu says. "It would demonstrate we're not getting fooled."

JWST has ushered in a new era of science, and despite the uncertainties, the rapid communication of new discoveries has invigorated astronomers. "It's been fantastic," Treu says. "It's really wonderful to see the community so engaged and excited." Now the question is, if we can truly believe what we are seeing, is it time to reappraise our understanding of the dawn of time? "We're peering into the unknown," Mason says. SA

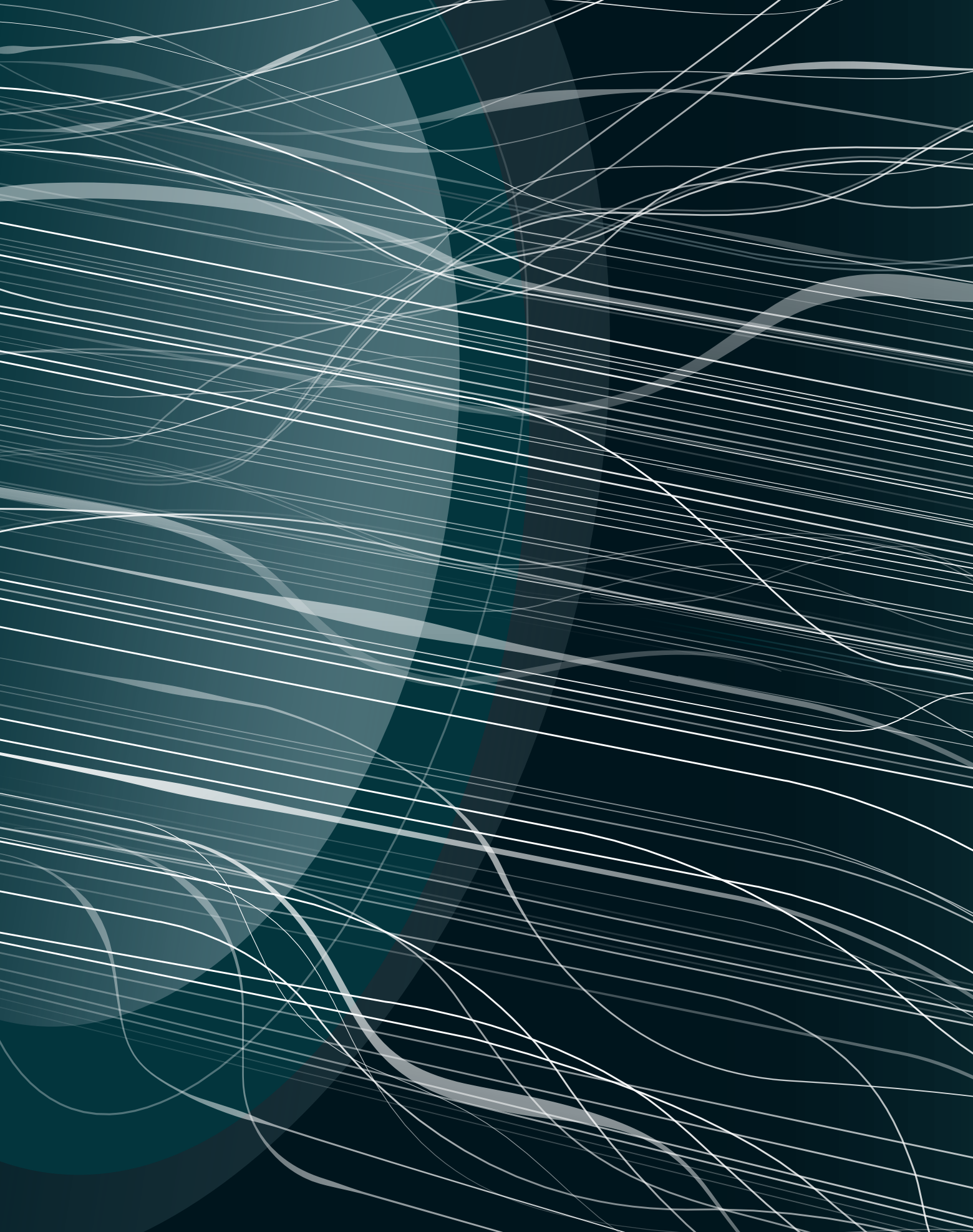
Jonathan O'Callaghan is a freelance journalist covering commercial spaceflight, space exploration and astrophysics.

NASA, ESA, CSA, STScI and Webb ERO Production Team

© 2023 Scientific American

THE CARTWHEEL GALAXY displays its characteristic dust-rich "spokes" and starry inner and outer rings in this near-infrared view from JWST. These features ripple like shock waves from the galaxy's center, the site of a high-speed collision with another galaxy some 400 million years ago.

© 2023 Scientific American



© 2023 Scientific American

Black Holes, Wormholes and Entanglement

Researchers cracked a paradox by considering what happens when the insides of black holes are connected by spacetime wormholes

By Ahmed Almheiri

Illustration by Harry Campbell

© 2023 Scientific American

THEORETICAL PHYSICS HAS BEEN IN CRISIS MODE EVER SINCE 1974, WHEN Stephen Hawking argued that black holes destroy information. Hawking showed that a black hole can evaporate, gradually transforming itself and anything it consumes into a featureless cloud of radiation. During the process, information about what fell into the black hole is apparently lost, violating a sacred principle of physics.

This remained an open problem for almost 50 years, but the pieces started falling into place in 2019 through research that I was involved in. The resolution is based on a new understanding of spacetime and how it can be rewired through quantum entanglement, which leads to the idea that part of the inside of a black hole, the so-called island, is secretly on the outside.

To understand how we arrived at these new ideas, we must begin with the inescapable nature of black holes.

A ONE-WAY STREET

NOTHING SEEMS MORE HOPELESS than trying to get out of a black hole—in fact, this impossibility is what defines black holes. They are formed when enough matter is confined within a small enough region that spacetime collapses in on itself in a violent feedback loop of squeezing and stretching that fuels more squeezing and stretching. These tidal forces run to infinity in finite time, marking the abrupt end of an entire region of spacetime at the so-called black hole singularity—the place where time stops and space ceases to make sense.

There is a fine line within the collapsing region that divides the area where escape is possible from the point of no return. This line is called the event horizon. It is the outermost point from which light barely avoids falling into the singularity. Unless a thing travels faster than light—a physical impossibility—it cannot escape from behind the event horizon; it is irretrievably stuck inside the black hole.

The one-way nature of this boundary is not immediately problematic. In fact, it is a robust prediction of the general theory of relativity. The danger starts when this theory interacts with the wild world of quantum mechanics.

SOMETHING OUT OF NOTHING

QUANTUM THEORY redeems black holes from being the greedy monsters they are made out to be. Every calorie of energy they consume they eventually give back in the form of Hawking radiation—energy squeezed out of the vacuum near the event horizon.

The idea of getting something out of nothing may sound absurd, but absurdity is not the worst allegation made against quantum mechanics. The emptiness of the vacuum in quantum theory belies a sea of particles—photons, electrons, gravitons, and more—that conspire to make empty space feel empty. These

© 2023 Scientific American

particles come in carefully arranged pairs, acting hand in hand as the glue that holds spacetime together.

Particle pairs that straddle the event horizon of a black hole, however, become forever separated from each other. The newly divorced particles peel away from the horizon in opposite directions, with one member crashing into the singularity and the other escaping the black hole's gravitational pull in the form of Hawking radiation. This process is draining for the black hole, causing it to get lighter and smaller as it emits energy in the form of the outgoing particles. Because of the law that energy must be conserved, the particles trapped inside must then carry negative energy to account for the decrease in the total energy of the black hole.

From the outside, the black hole appears to be burning away (although it happens so slowly, you can't see it happening in real life). When you burn a book, the words on its pages imprint themselves on the pattern of the emanating light and the remaining ashes. This information is thus preserved, at least in principle. If an evaporating black hole were a normal system like the burning book, then the information about what falls into it would be encoded into the emerging Hawking radiation. Unfortunately, this is complicated by the quantum-mechanical relation among the particles across the horizon.

EINSTEIN'S ENEMY

THE ISSUE BEGINS WITH THE END of the pairing of the two particles straddling the event horizon. Despite being separated, they maintain a quantum union that transcends space and time—they are connected by entanglement. Rejected as an absurdity by the physicists who predicted it, quantum entanglement is perhaps one of the weirdest aspects of our universe and arguably one of its most essential. The concept was first concocted by Albert Einstein, Boris Podolsky and Nathan Rosen as a rebuttal against what was then the nascent theory of quantum mechanics. They cited entanglement as a reason the theory must be incomplete—"spooky" is how Einstein famously described the phenomenon.

For a simple example of entanglement, consider two coins in a superposition—the quantum phenomenon of being in multiple states until a measurement is made—of both coins being either heads or tails. The coins aren't facing heads and tails at the same time—that's physically impossible—but the superposition signifies that the chance of observing the pair of coins in either orientation, both heads or both tails, is a probability of one half. There is no chance of ever finding the coins in opposite orientations. The two coins are entangled; the measurement result of one predicts the result of the other with complete certainty. Either coin by itself is completely random, devoid of information, but the randomness of the pair is perfectly correlated.

The scientists were troubled by how the two coins appeared to influence each other without having to come into physical contact. The coins could be in separate galaxies while still maintaining the same amount of entanglement between them. Einstein was unnerved by the apparent "spooky action at a distance" linking the results of the two separate random measurements.

The irony is that Einstein himself is in a superposition of being both wrong and right. He was right to recognize the importance of entanglement in distinguishing quantum mechanics from classical physics. What he got wrong can be summed up with the truism "correlation does not imply causation." Although the fates of the particles are inextricably correlated, the measurement outcome of one does not *cause* the outcome of the other. It turns out that quantum mechanics simply allows for a new, higher degree of correlation than we are used to.

THE IDEA OF GETTING SOMETHING OUT OF NOTHING MAY SOUND ABSURD, BUT ABSURDITY IS NOT THE WORST ALLEGATION MADE AGAINST QUANTUM MECHANICS.

INFORMATION LOST

BECAUSE HAWKING RADIATION is composed of one half of a collection of entangled pairs, it emerges from the black hole in a completely random state—if the particles were coins, they would be observed to be heads or tails with equal probability. Hence, we cannot infer anything useful about the contents of the black hole from the random measurements of the radiation. This means that an evaporating black hole is basically a glorified information shredder, except unlike the mechanical kind, it does a thorough job.

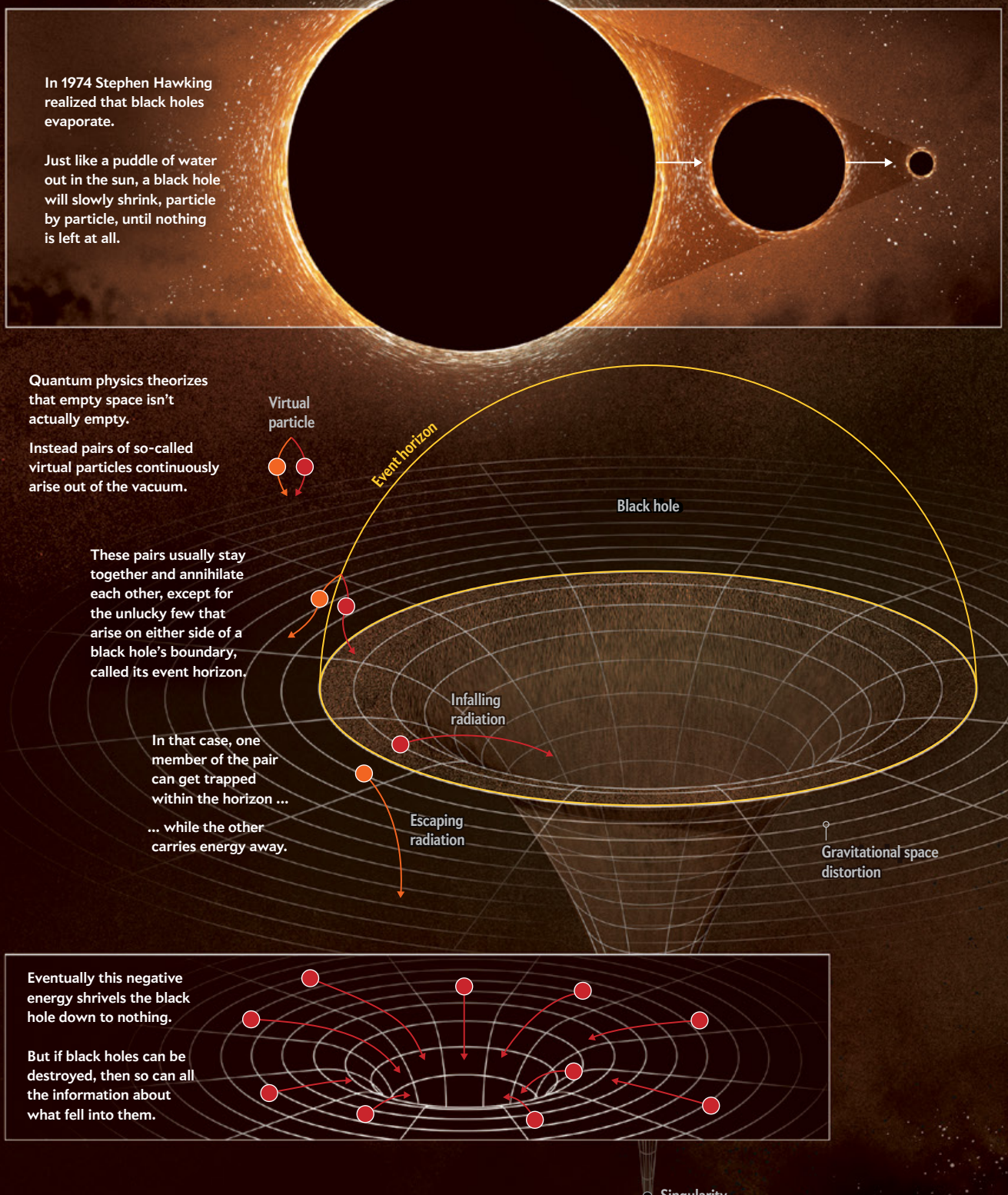
We can measure the lack of information—or the randomness—in the Hawking radiation by thinking about the amount of entanglement between the radiation and the black hole. This is because one member of an entangled pair is always random, and the outside members are all that remains by the end of the evaporation. The calculation of randomness goes by many names, including entanglement entropy, and it grows with every emerging Hawking particle, plateauing at a large value once the black hole has completely disappeared.

This pattern differs from what happens when information is preserved, as in the example of a burning book. In such a case, the entropy may rise initially, but it has to peak and then fall to zero by the end of the process. The intuition behind this rule is clear

© 2023 Scientific American

The Black Hole Information Paradox

For almost half a century physicists have agonized over the question of what happens to the stuff that falls into a black hole. If, as theory predicts, black holes destroy information about everything that has ever fallen in, then our theories of nature are in deep, foundational trouble. In recent years, though, scientists have made a breakthrough that might finally solve the puzzle.


In 1974 Stephen Hawking realized that black holes evaporate.
Just like a puddle of water out in the sun, a black hole will slowly shrink, particle by particle, until nothing is left at all.
Quantum physics theorizes that empty space isn't actually empty.
Instead pairs of so-called virtual particles continuously arise out of the vacuum.
Virtual particle
Event horizon
Black hole
These pairs usually stay together and annihilate each other, except for the unlucky few that arise on either side of a black hole's boundary, called its event horizon.
Infalling radiation
In that case, one member of the pair can get trapped within the horizon ...
... while the other carries energy away.
Escaping radiation
Gravitational space distortion
Eventually this negative energy shrivels the black hole down to nothing.
But if black holes can be destroyed, then so can all the information about what fell into them.
Singularity

Graphic by Matthew Twombly

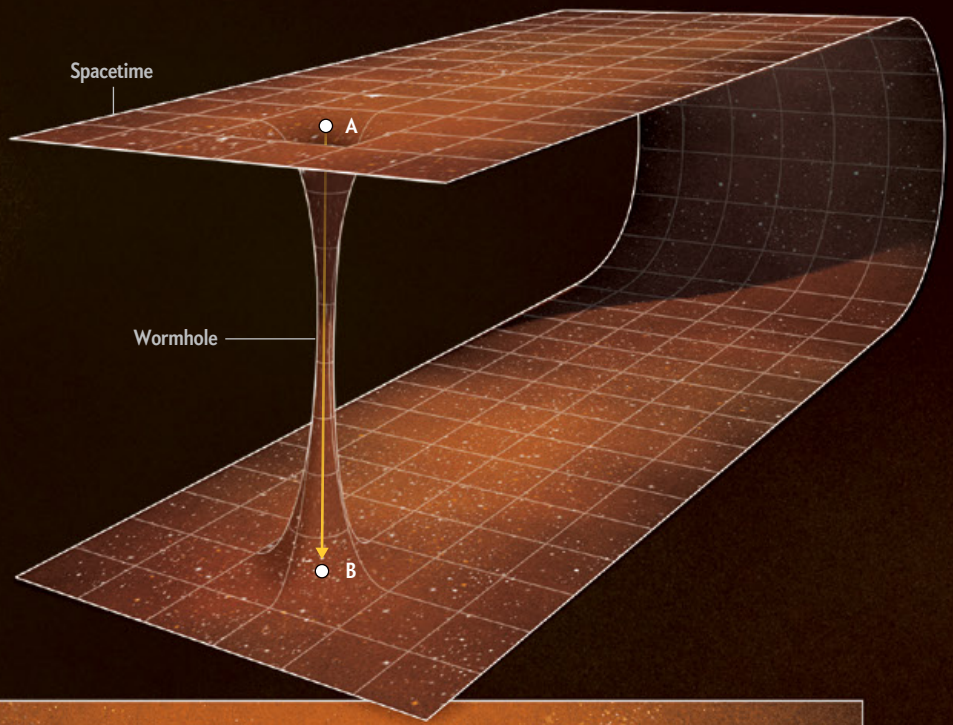
© 2023 Scientific American

That seems to break a fundamental law of physics, which says that information can never be destroyed.

This is the paradox.

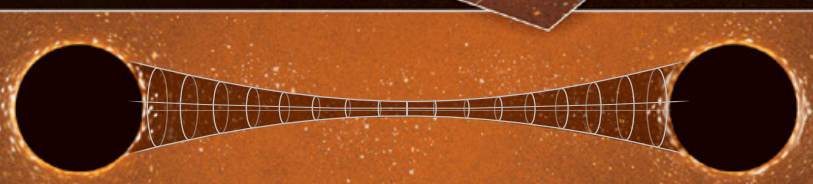
In the past few years a unique solution has revealed itself: wormholes.

Wormholes are theoretical bridges in spacetime that connect two distant spots through a shortcut.



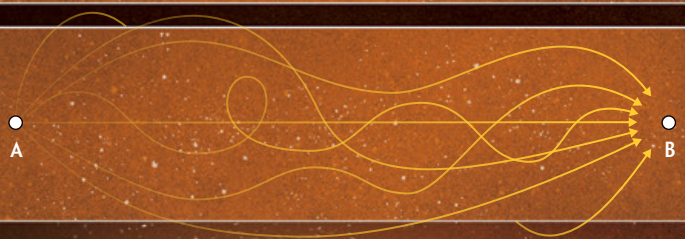
The inside of a black hole could be connected to the inside of another black hole via a wormhole.

Though rare, it's theoretically possible.



And in quantum physics, everything that can happen does happen.

A particle doesn't simply travel along one particular path from point A to point B.



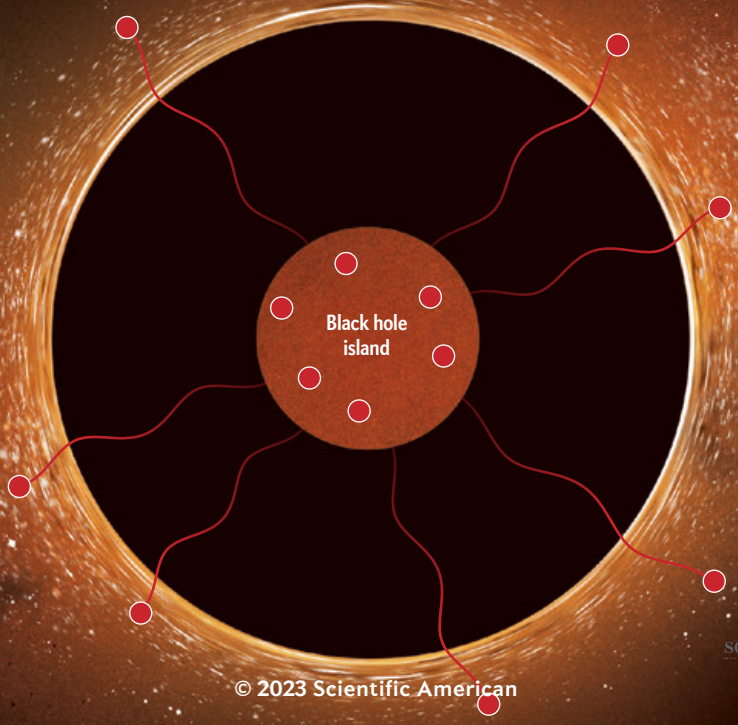
It takes all of them simultaneously.

If wormholes are at the center of black holes, information pulled within may not be destroyed.

Instead the interiors of black holes seem to contain special areas deep inside called islands.

These islands are both inside and outside the black holes, as if they are part of the escaping radiation that is depleting the black holes over time.

And as they escape, the information within them escapes, too.



These new ideas are pretty confounding, even to physicists, who are discovering that the cosmos and the nature of our reality are even weirder than we could have ever imagined.

© 2023 Scientific American

when you think about a standard deck of cards: suppose you are dealt cards from a 52-card deck, one by one, facing down. The entropy of the cards in your possession is simply a measure of your ignorance of what's on the other side of the cards—specifically, the number of possibilities of what they could be. If you have been dealt just one card, the entropy is 52 because there are 52 possibilities. But as you are dealt more, the entropy rises, peaking at 500 trillion for 26 cards, which could be any of 500 trillion different combinations. After this, though, the possible mixes of cards, and thus the entropy, go back down, reaching 52 again when you have 51 cards. Once you have all the cards, you are certain of exactly what you have—the entire deck—and the entropy is zero. This rising and falling pattern of entropy, known as the Page curve, applies to all normal quantum-mechanical systems. The time at which the entropy peaks and then starts to decrease is the Page time.

EVENTUALLY MY COLLEAGUES AND I REALIZED THAT BOTH THE INFORMATION PARADOX AND THE NEWER FIREWALL PARADOX AROSE BECAUSE OUR ATTEMPTS TO MELD QUANTUM MECHANICS AND BLACK HOLE PHYSICS WERE TOO TIMID.

The destruction of information inside black holes spells disaster for physics because the laws of quantum mechanics stipulate that information cannot be obliterated. This is the famous information paradox—the fact that a sprinkling of quantum mechanics onto the description of black holes leads to a seemingly insurmountable inconsistency. Physicists knew we needed a more complete understanding of quantum-gravitational physics to generate the Page curve for the Hawking radiation. Unsurprisingly, this task proved difficult.

AN EVENTFUL HORIZON

PART OF THE CHALLENGE was that no minor tweaking of the evaporation process was sufficient to generate the Page curve and send the entropy back down to zero. What we needed was a drastic reimagining of the structure of a black hole.

In a paper I published in 2013 with Donald Marolf, the late Joseph Polchinski, and Jamie Sully (known collectively as AMPS), we tried out several ways to modify the picture of evaporating black holes using a series of gedankenexperiments—the German term for the kind of thought experiments Einstein popularized. Through our trials we concluded that to save the sanctity of information, one of two things had to give: either physics must be nonlocal—allowing for information to instantaneously disappear from the interior and appear outside the event horizon—or a new process must kick in at the Page time. To preclude the increase of entropy, this process would have to break the entanglement between the particle pairs across the event horizon. The former option—making physics nonlocal—was too radical, so we decided to go with the latter.

This modification helps to preserve information, but it poses another paradox. Recall that the entanglement across the horizon was a result of having empty space there—the way the vacuum is maintained by a sea of entangled pairs of particles. The entanglement is key; breaking it comes at the cost of creating a wall of extremely high-energy particles, which our group named the firewall. Having such a firewall at the horizon would forbid anything from entering the black hole. Instead infalling matter would be vaporized on contact. The black hole at the Page time would suddenly lose its interior, and spacetime would come to an end, not at the singularity deep inside the black hole but right there at the event horizon. This conclusion is known as the firewall paradox, a catch-22 that meant any solution to the information paradox must come at the cost of destroying what we know about black holes. If ever there were a quagmire, this would be it.

FLUCTUATING WORMHOLES

EVENTUALLY MY COLLEAGUES and I realized that both the information paradox and the newer firewall paradox arose because our attempts to meld quantum mechanics and black hole physics were too timid. It wasn't enough to apply quantum mechanics to only the matter present in black holes—we had to devise a quantum treatment of the black hole spacetime as well. Although quantum effects on spacetime are usually very small, they could be enhanced by the large entanglement produced by the evaporation. Such an effect may be subtle, but its implications would be huge.

To consider the quantum nature of spacetime, we relied on a technique designed by Richard Feynman called the path integral of quantum mechanics. The idea is based on the weird truth that, according to quantum theory, particles don't simply travel along a single path from point A to point B—they travel along *all* the different paths connecting the two points. The path integral is a way of describing a particle's travels in terms of a quantum superposition of all possible routes. Similarly, a quantum spacetime can be in a superposition of different complicated shapes evolving in different ways. For instance, if we start and end with two regular black holes, the quantum spacetime within them has a nonzero probabil-

© 2023 Scientific American

ity of creating a short-lived wormhole that temporarily bridges their interiors.

Usually the probability of this happening is vanishingly slim. When we carry out the path integral in the presence of the Hawking radiation of multiple black holes, however, the large entanglement between the Hawking radiation and the black hole interiors amplifies the likelihood of such wormholes. This realization came to me through work I did in 2019 with Thomas Hartman, Juan Maldacena, Edgar Shaghoulian and Amirhossein Tajdini, and it was also the result of an independent collaboration by Geoffrey Penington, Stephen Shenker, Douglas Stanford and Zhenbin Yang.

ISLANDS BEYOND THE HORIZON

WHY DOES IT MATTER if some black holes are connected by wormholes? It turns out that they modify the answer of how much entanglement entropy there is between the black hole and its Hawking radiation. The key is to measure this entanglement entropy in the presence of multiple copies of the system. This is known as the replica trick.

The relevant physical effect of these temporary wormholes is to swap out the interiors among the different black holes. This happens literally: what was in one black hole gets shoved into one of the other copies far away, and the original black hole assumes a new spacetime interior from a different one. The swapped region of the black hole interior is called the island, and it encompasses almost the entire interior up to the event horizon.

The swapping is exactly what the doctor ordered! Focusing on one of the black holes and its Hawking radiation, the swapped-out island takes with it all the partner particles that are entangled with the outgoing Hawking radiation, and hence, technically, there is no entanglement between the black hole and its radiation.

Including this potential effect of wormholes produces a new formula for the entanglement entropy of the radiation when applied to a single copy of the system. Instead of Hawking's original calculation, which simply counts the number of Hawking particles outside a black hole, the new formula curiously treats the island as if it were outside and a part of the exterior Hawking radiation. Therefore, the entanglement between the island and the exterior should not be counted toward the entropy. Instead the entropy that it predicts comes almost entirely from the probability of the swap actually occurring, which is equal to the area of the boundary of the island—roughly the area of the event horizon—divided by Newton's gravitational constant. As the black hole shrinks, this contribution to the entropy decreases. This is the island formula for the entanglement entropy of the Hawking radiation.

The final step in computing the entropy is to take the minimum between the island formula and Hawking's original calculation. This gives us the Page curve that we've been after. Initially we calculate the entanglement entropy of the radiation with Hawking's original formula because the answer starts off smaller than the area of the event horizon of the black hole. But as the black hole evaporates, the area shrinks, and the new formula takes the baton as the true representative of the radiation's entanglement entropy.

What is remarkable about this result is that it solves two paradoxes with one formula. It appears to address the firewall paradox by supporting the option of nonlocality that my AMPS group originally dismissed. Instead of breaking the entanglement at the horizon, we are instructed to treat the inside—the island—as part of the outside. The island itself becomes nonlocally mapped to the outside. And the formula solves the information paradox by revealing how black holes produce the Page curve and preserve information.

Let's take a step back and think about how we got here. The origins of the information paradox can be traced back to the incompatibility between the sequestering of information by the event horizon and the quantum-mechanical requirement of information flow outside the black hole. Naive resolutions of this tension lead to drastic modifications of the structure of black holes; however, subtle yet dramatic effects from fluctuating wormholes change everything. What emerges is a self-consistent picture that lets a black hole retain its regular structure as predicted by general relativity, albeit with the presence of an implicit though powerful nonlocality. This nonlocality is screaming that we should consider a portion of the black hole's interior—the island—as part of the exterior, as a single unit with the outside radiation. Thus, information can escape a black hole not by surmounting the insurmountable event horizon but by simply falling deeper into the island.

Despite the excitement of this breakthrough, we have only begun to explore the implications of spacetime wormholes and the island formula. Curiously, while they ensure that the island is mapped onto the radiation, they do not generate a definite prediction for specific measurements of the Hawking radiation. What they do teach us, however, is that wormholes are the missing ingredient in Hawking's original estimation of the randomness in the radiation and that gravity is in fact smart enough to comply with quantum mechanics.

Through these wormholes, gravity harnesses the power of entanglement to achieve nonlocality, which is just as unnerving to us as the entanglement that originally spooked Einstein. We must admit that, at some level, Einstein was right after all. SA

Ahmed Almheiri is a theoretical physicist at New York University in Abu Dhabi, United Arab Emirates, where he studies the connections between quantum information and quantum gravity.

© 2023 Scientific American

The Universe Is Not Locally Real

© 2023 Scientific American

Experiments with entangled light have revealed a profound mystery at the heart of reality

By Daniel Garisto

© 2023 Scientific American

ONE OF THE MORE UNSETTLING DISCOVERIES IN THE PAST HALF A century is that the universe is not locally real. In this context, "real" means that objects have definite properties independent of observation—an apple can be red even when no one is looking. "Local" means that objects can be influenced only by their surroundings and that any influence cannot travel faster than light. Investigations at the frontiers of quantum physics have found that these things cannot both be true. Instead the evidence shows that objects are *not* influenced solely by their surroundings, and they *may* also lack definite properties prior to measurement.

This is, of course, deeply contrary to our everyday experiences. As Albert Einstein once bemoaned to a friend, "Do you really believe the moon is not there when you are not looking at it?" To adapt a phrase from author Douglas Adams, the demise of local realism has made a lot of people very angry and has been widely regarded as a bad move.

Blame for this achievement has now been laid squarely on the shoulders of three physicists: John Clauser, Alain Aspect and Anton Zeilinger. They equally split the 2022 Nobel Prize in Physics "for experiments with entangled photons, establishing the violation of Bell inequalities and pioneering quantum information science." ("Bell inequalities" refers to the pioneering work of Northern Ireland physicist John Stewart Bell, who laid the foundations for the 2022 Physics Nobel in the early 1960s.) Colleagues agreed that the trio had it coming, deserving this reckoning for overthrowing reality as we know it. "It was long overdue," says Sandu Popescu, a quantum physicist at the University of Bristol in England. "Without any doubt, the prize is well deserved."

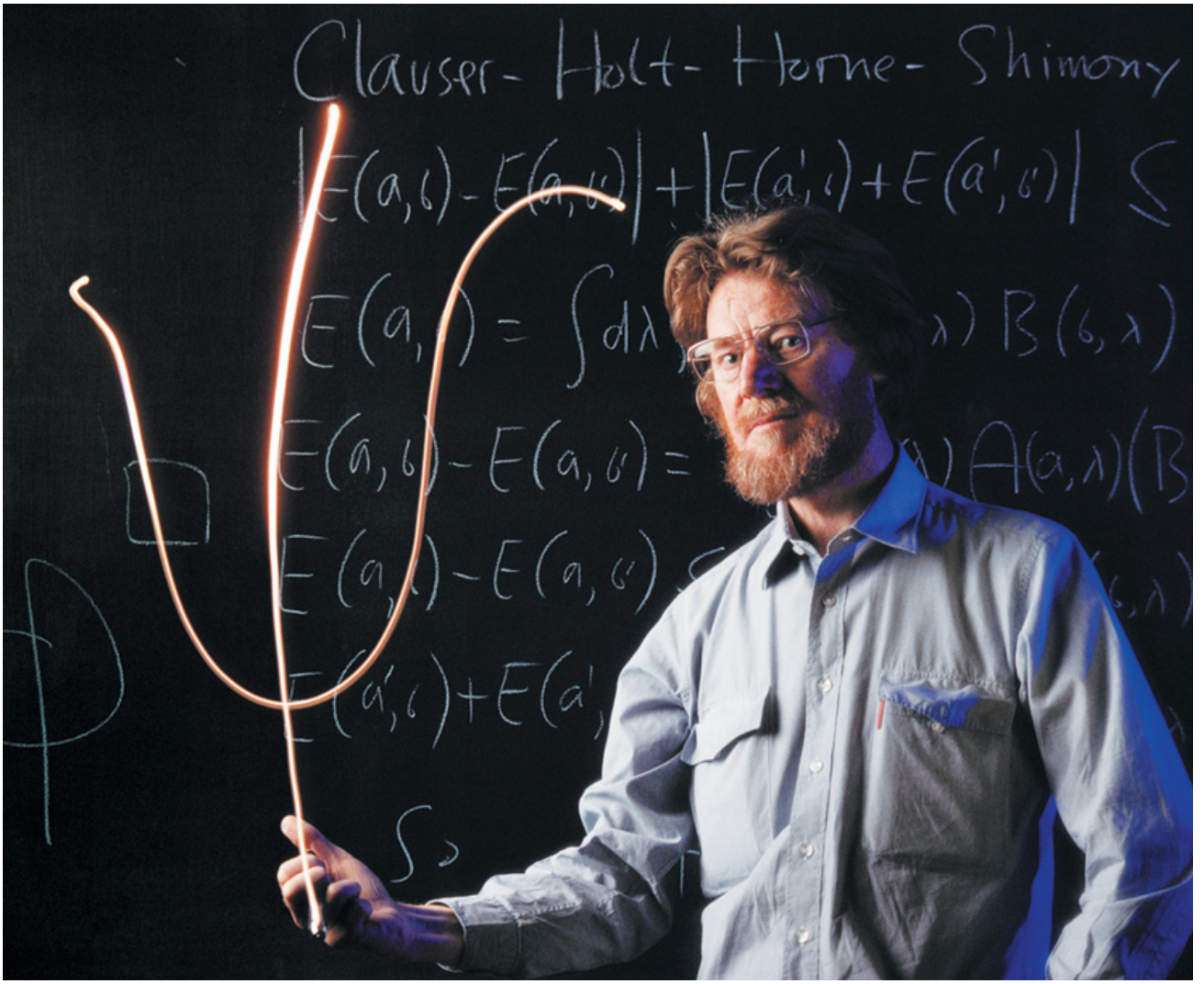
"The experiments beginning with the earliest one of Clauser and continuing along show that this stuff isn't just philosophical, it's real—and like other real things, potentially useful," says Charles H. Bennett, an eminent quantum researcher at IBM. "Each year I thought, 'Oh, maybe this is the year,'" says David Kaiser, a physicist and historian at the Massachusetts Institute of Technology. In 2022 "it really was. It was very emotional—and very thrilling."

The journey from fringe to favor was a long one. From about 1940 until as late as 1990, studies of so-called quantum foundations were often treated as philosophy at best and crackpottery at worst. Many scientific journals refused to publish papers on the topic, and academic positions indulging such investigations were nearly impossible to come by. In 1985 Popescu's adviser warned him against a Ph.D. in the subject. "He said, 'Look, if you do that, you will have fun for five years, and then you will be jobless,'" Popescu says.

Today quantum information science is among the most vibrant subfields in all of physics. It links Einstein's

Athul Satheesh/500px/Getty Images (*preceding pages*)

© 2023 Scientific American



Peter Menzel/Science Source

THE WORK of John Stewart Bell in the 1960s sparked a quiet revolution in quantum physics.

general theory of relativity with quantum mechanics via the still mysterious behavior of black holes. It dictates the design and function of quantum sensors, which are increasingly being used to study everything from earthquakes to dark matter. And it clarifies the often confusing nature of quantum entanglement, a phenomenon that is pivotal to modern materials science and that lies at the heart of quantum computing.

"What even makes a quantum computer 'quantum?'" Nicole Yunger Halpern, a National Institute of Standards and Technology physicist, asks rhetorically. "One of the most popular answers is entanglement, and the main reason we understand entanglement is the grand work participated in by Bell and these Nobel Prize winners. Without that understanding of entanglement, we probably wouldn't be able to realize quantum computers."

FOR WHOM THE BELL TOLLS

THE TROUBLE WITH quantum mechanics was never that it made the wrong predictions—in fact, the theory described the microscopic world splendidly right from the start when physicists devised it in the opening decades of the 20th century. What Einstein, Boris Podolsky and Nathan Rosen took issue with, as they explained in their iconic 1935 paper, was the theory's uncomfortable implications for reality. Their analysis, known by their initials EPR, centered on a thought experiment meant to illustrate the absurdity of quantum mechanics. The goal was to show how under certain conditions the theory can break—or at least deliver nonsensical results that conflict with our deepest assumptions about reality.

A simplified and modernized version of EPR goes something like this: Pairs of particles are sent off in different directions from a common source, targeted for two observers, Alice and Bob, stationed at opposite ends of the solar system. Quantum mechanics dictates that it is impossible to know the spin, a quantum property of individual particles, prior to measurement. Once Alice measures one of her particles, she finds its spin to be either "up" or "down." Her results are random, and yet when she measures up, she instantly knows that Bob's corresponding particle—

© 2023 Scientific American

which had a random, indefinite spin—must now be down. At first glance, this is not so odd. Maybe the particles are like a pair of socks—if Alice gets the right sock, Bob must have the left.

But under quantum mechanics, particles are not like socks, and only when measured do they settle on a spin of up or down. This is EPR's key conundrum: If Alice's particles lack a spin until measurement, then how (as they whiz past Neptune) do they know what Bob's particles will do as they fly out of the solar system in the other direction? Each time Alice measures, she quizzes her particle on what Bob will get if he flips a coin: up or down? The odds of correctly predicting this even 200 times in a row are one in 10^{60}—a number greater than the total number of atoms in the solar system. Yet despite the billions of kilometers that separate the paired particles, quantum mechanics says Alice's particles can keep correctly predicting, as though they were telepathically connected to Bob's particles.

Designed to reveal the incompleteness of quantum mechanics, EPR eventually led to experimental results that instead reinforce the theory's most mind-boggling tenets. Under quantum mechanics, nature is not locally real: particles may lack properties such as spin up or spin down prior to measurement, and they seem to talk to one another no matter the distance. (Because the outcomes of measurements are random, these correlations cannot be used for faster-than-light communication.)

Physicists skeptical of quantum mechanics proposed that this puzzle could be explained by hidden variables, factors that existed in some imperceptible level of reality, beneath the subatomic realm, that contained information about a particle's future state. They hoped that in hidden variable theories, nature could recover the local realism denied it by quantum mechanics. "One would have thought that the arguments of Einstein, Podolsky and Rosen would produce a revolution at that moment, and everybody would have started working on hidden variables," Popescu says.

Einstein's "attack" on quantum mechanics, however, did not catch on among physicists, who by and large accepted quantum mechanics as is. This was less a thoughtful embrace of nonlocal reality than a desire not to think too hard—a head-in-the-sand sentiment later summarized by American physicist N. David Mermin as a demand to "shut up and calculate." The lack of interest persisted in part because John von Neumann, a highly regarded scientist, had in 1932 published a mathematical proof ruling out hidden variable theories. Von Neumann's proof, it must be said, was refuted just three years later by a young female mathematician, Grete Hermann, but at the time no one seemed to notice.

The problem of nonlocal realism would languish for another three decades before being shattered by Bell. From the start of his career, Bell was bothered by quantum orthodoxy and was sympathetic toward hidden variable theories. Inspiration struck him in 1952, when he learned that American physicist David Bohm had formulated a viable nonlocal hidden variable interpretation of quantum mechanics—something von Neumann had claimed was impossible.

Bell mulled the ideas for years, as a side project to his job working as a particle physicist at CERN near Geneva. In 1964 he recognized the same flaws in von Neumann's argument that Hermann had discovered. And then, in a triumph of rigorous thinking, Bell concocted a theorem that dragged the question of local hidden variables from its metaphysical quagmire onto the concrete ground of experiment.

Typically local hidden variable theories and quantum mechanics predict indistinguishable experimental outcomes. What Bell realized is that under precise circumstances, an empirical discrepancy between the two can emerge. In the eponymous Bell test (an evolution of the EPR thought experiment), Alice and Bob receive the same paired particles, but now they each have two different detector settings—A and a, B and b. These detector settings are an additional trick to throw off Alice and Bob's apparent telepathy. In local hidden variable theories, one particle cannot know which question the other is asked. Their correlation is secretly set ahead of time and is not sensitive to updated detector settings. But according to quantum mechanics, when Alice and Bob use the same settings (both uppercase or both lowercase), each particle is aware of the question the other is posed, and the two will correlate perfectly—in sync in a way no local theory can account for. They are, in a word, entangled.

Measuring the correlation multiple times for many particle pairs, therefore, could prove which theory was correct. If the correlation remained below a limit derived from Bell's theorem, this would suggest hidden variables were real; if it exceeded Bell's limit, then the mind-boggling tenets of quantum mechanics would reign supreme. And yet, in spite of its potential to help determine the nature of reality, Bell's theorem languished unnoticed in a relatively obscure journal for years.

THE BELL TOLLS FOR THEE

IN 1967 A GRADUATE STUDENT at Columbia University named John Clauser accidentally stumbled across a library copy of Bell's paper and became enthralled by the possibility of proving hidden variable theories correct. When Clauser wrote to Bell two years later, asking if anyone had performed the test, it was among the first feedback Bell had received.

Three years after that, with Bell's encouragement, Clauser and his graduate student Stuart Freedman performed the first Bell test. Clauser had secured permission from his supervisors but little in the way of funds, so he became, as he said in a later interview, adept at "dumpster diving" to secure equipment—some of which he and Freedman then duct-taped together. In Clauser's setup—a kayak-sized apparatus requiring careful tuning by hand—pairs of photons were sent in

© 2023 Scientific American

opposite directions toward detectors that could measure their state, or polarization.

Unfortunately for Clauser and his infatuation with hidden variables, once he and Freedman completed their analysis, they had to conclude that they had found strong evidence against them. Still, the result was hardly conclusive because of various "loopholes" in the experiment that conceivably could allow the influence of hidden variables to slip through undetected. The most concerning of these was the locality loophole: if either the photon source or the detectors could have somehow shared information (which was plausible within an object the size of a kayak), the resulting measured correlations could still emerge from hidden variables. As David Kaiser explained, if Alice tweets at Bob to tell him her detector setting, that interference makes ruling out hidden variables impossible.

Closing the locality loophole is easier said than done. The detector setting must be quickly changed while photons are on the fly—"quickly" meaning in a matter of mere nanoseconds. In 1976 a young French expert in optics, Alain Aspect, proposed a way for doing this ultra-speedy switch. His group's experimental results, published in 1982, only bolstered Clauser's results: local hidden variables looked extremely unlikely. "Perhaps Nature is not so queer as quantum mechanics," Bell wrote in response to Aspect's test. "But the experimental situation is not very encouraging from this point of view."

Other loopholes remained, however, and Bell died in 1990 without witnessing their closure. Even Aspect's experiment hadn't fully ruled out local effects, because it took place over too small a distance. Similarly, as Clauser and others had realized, if Alice and Bob detected an unrepresentative sample of particles—like a survey that contacted only right-handed people—their experiments could reach the wrong conclusions.

No one pounced to close these loopholes with more gusto than Anton Zeilinger, an ambitious, gregarious Austrian physicist. In 1997 he and his team improved on Aspect's earlier work by conducting a Bell test over a then unprecedented distance of nearly half a kilometer. The era of divining reality's nonlocality from kayak-sized experiments had drawn to a close. Finally, in 2013, Zeilinger's group took the next logical step, tackling multiple loopholes at the same time.

"Before quantum mechanics, I actually was interested in engineering. I like building things with my hands," says Marissa Giustina, a quantum researcher at Google who worked with Zeilinger. "In retrospect, a loophole-free Bell experiment is a giant systems-engineering project." One requirement for creating an experiment closing multiple loopholes was finding a perfectly straight, unoccupied 60-meter tunnel with access to fiber-optic cables. As it turned out, the dungeon of Vienna's Hofburg palace was an almost ideal setting—aside from being caked with a century's worth of dust. Their results, published in 2015, coincided with similar tests from two other groups that also found quantum mechanics as flawless as ever.

BELL'S TEST REACHES THE STARS

ONE GREAT FINAL LOOPHOLE remained to be closed—or at least narrowed. Any prior physical connection between components, no matter how distant in the past, has the possibility of interfering with the validity of a Bell test's results. If Alice shakes Bob's hand prior to departing on a spaceship, they share a past. It is seemingly implausible that a local hidden variable theory would exploit these loopholes, but it was still possible.

In 2016 a team that included Kaiser and Zeilinger performed a cosmic Bell test. Using telescopes in the Canary Islands, the researchers sourced random decisions for detector settings from stars sufficiently far apart in the sky that light from one would not reach the other for hundreds of years, ensuring a centuries-spanning gap in their shared cosmic past. Yet even then, quantum mechanics again proved triumphant.

Today quantum information science is among the most vibrant subfields in all of physics.

One of the principal difficulties in explaining the importance of Bell tests to the public—as well as to skeptical physicists—is the perception that the veracity of quantum mechanics was a foregone conclusion. After all, researchers have measured many key aspects of quantum mechanics to a precision of greater than 10 parts in a billion. "I actually didn't want to work on it," Giustina says. "I thought, like, 'Come on, this is old physics. We all know what's going to happen.'" But the accuracy of quantum mechanics could not rule out the possibility of local hidden variables; only Bell tests could do that.

"What drew each of these Nobel recipients to the topic, and what drew John Bell himself to the topic, was indeed [the question], 'Can the world work that way?'" Kaiser says. "And how do we really know with confidence?" What Bell tests allow physicists to do is remove the bias of anthropocentric aesthetic judgments from the equation. They purge from the work the parts of human cognition that recoil at the possibility of eerily inexplicable entanglement or that scoff at hidden variable theories as just more debates over how many angels may dance on the head of a pin.

The award honors Clauser, Aspect and Zeilinger, but it is testament to all the researchers who were unsatisfied with superficial explanations about quantum mechanics and who asked their questions even when doing so was unpopular. "Bell tests," Giustina concludes, "are a very useful way of looking at reality." SA

Daniel Garisto is a freelance science journalist covering advances in physics and other natural sciences. He is based in New York.

© 2023 Scientific American

THE HUMAN ENGINE

Studies of metabolism reveal surprising insights into how we burn calories—and how cooperative food production helped *Homo sapiens* flourish

By Herman Pontzer

Illustration by Eva Vázquez

© 2023 Scientific American

© 2023 Scientific American

IT WAS MY DAUGHTER CLARA'S SEVENTH BIRTHDAY PARTY, A SCENE AT ONCE FAMILIAR and bizarre. The celebration was an American take on a classic script: a shared meal of pizza and picnic food, a few close COVID-compliant friends and family, a beaming kid blowing out candles on a heavily iced cake. With roughly 380,000 boys and girls around the world turning seven each day, it was a ritual no doubt repeated by many, the world's most prolific primate singing "Happy Birthday" in an unbroken global chorus.

Such a wholesome setting seems an unlikely place for rampant rule breaking. But as an evolutionary anthropologist, I can't help but notice the blatant disregard our species shows for the natural order. Nearly every aspect of our modern lives marks a cheerfully outrageous departure from the laws that govern every other species on the planet, and this birthday party was no exception. Aside from the fresh veggies left wilting in the sun, none of the food was recognizable as a product of nature. The cake was a heat-treated amalgam of pulverized grass seed, chicken eggs, cow milk and extracted beet sugar. The raw materials for the snacks and drinks would take a forensic chemist years to reconstruct. It was a calorie bonanza that animals foraging in the wild could only dream about, and we were giving it away to people who didn't even share our genes. All this to celebrate some obscure astronomical alignment, the moment our planet swept through the same position relative to its star as on the day my daughter was born. At seven years old, most mammals are grandparents if they're lucky enough to be alive. Clara was still a kid, dependent on us for food and shelter and years away from independence.

Humans weren't always such scofflaws. We come from a good Family. The living apes, our closest relatives, are well-behaved primates, eating fruit and leaves straight from the tree and nibbling on the occasional meal of insects or small game. Like every other mammal, apes learn early to fend for themselves, foraging on their own as soon as they're weaned, and they know better than to give their hard-earned food away. Fossils from deep in the human lineage, the first four million years after we broke from the other apes, indicate our early ancestors played by the same ecological rules.

Around 2.5 million years ago things took an unlikely turn. Early populations of the genus *Homo* stumbled onto a new way of making a living, something unprecedented in the history of life. Instead of pursuing a career as a plant eater, carnivore or generalist, they tried a strange, dual strategy: some would hunt, others would gather, and they'd share whatever they acquired. This cooperative approach placed a premium on intelligence, and over millennia brain size began to increase. Our Paleolithic ancestors learned to knap delicate blades from round stone cobbles, hunt large game and cook their food. They built hearths and homes and began changing the landscape, developing an ecological mastery that led eventually to farming.

These evolutionary shifts reverberate today. The cooperative foraging that pushed our hunting, gathering and farming ancestors to flout long-established ecological rules didn't just change the foods we eat. It altered fundamental aspects of our biology, including our metabolism. The same unlikely series of events that gave us birthday cake has also shaped the way we eat it—and how we use the calories.

For all the talk about metabolism in the exercise and dieting worlds, you would think the science was settled. In reality, we've been embarrassingly short on hard data about the calories we burn each day and how we evolved to obtain them. But recently my colleagues and I have made important strides in understanding how our bodies use energy. Our findings have overturned much of the received wisdom about the ways human energy requirements change over the course of a lifetime. And, as we've discovered in a parallel effort, our energy needs are deeply intertwined with the evolution of our food-production strategies: foraging and farming. Together these studies provide the clearest picture yet of the inner workings of the human engine—and how our strategy for earning, burning and sharing calories underpins our extraordinary success as a species.

ENERGY BUDGETS

OUR BODIES are wonders of coordinated chaos. Every second of every day, each of your 37 trillion cells is hard at work, pulling in nutrients, building new proteins and doing the myriad other tasks that keep you alive. All of this work takes energy. Our metabolism is the energy we expend (or the calories we burn) each day. That energy comes from the food we eat, and so our metabolism also sets our energy requirements. Calories in, calories out.

Evolutionary biologists often think about metabolism as an organism's energy budget. Life's essential tasks, including growth, reproduction and bodily maintenance, require energy. And every organism must balance its books.

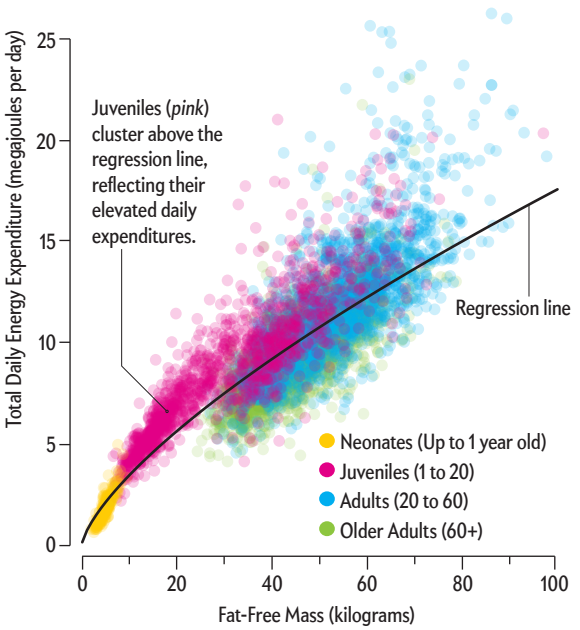
Humans are a striking example of this evolutionary bookkeeping in action. The traits that distinguish us from the other apes, including our huge brains, big babies and long lives, all require a lot of energy. We pay for some of these costs by spending less on

© 2023 Scientific American

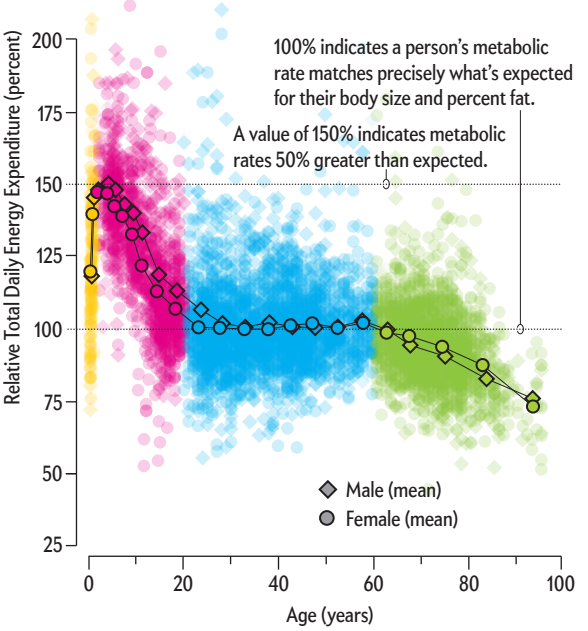
Measuring Metabolism

The first comprehensive study investigating the effects of age and body size on daily energy expenditure has upended much of the conventional wisdom about metabolism. Metabolic rates increase with body size, as expected. But they are not inherently different in men and women, nor do they decline with middle age, among other revelations from this research.

The greatest predictor of metabolism is fat-free mass. Larger bodies generally burn more calories.



Metabolism skyrockets over the first year of life. Daily energy expenditures hold remarkably steady from age 20 to 60.



our digestive system, having evolved a shorter intestinal tract and smaller liver. But we have also increased our metabolic rate and the size of our energy budget. For our body size, humans consume and burn more calories each day than any of the other apes. Our cells have evolved to work harder.

The work our bodies do changes as we age, the activities of our cells waxing and waning in a choreographed dance from growth to adulthood to senescence. Tracking those changes through our metabolism could provide a better understanding of the work our cells do at each age as well as our changing calorie needs. But a clear audit of our metabolism over the human life span has been hard to obtain.

It's obvious that adults need more calories than infants—bigger people have more cells doing more work, so they burn more energy. We also know that elderly people tend to eat less, although that's often accompanied by a loss of body weight, particularly muscle mass. But if we want to know how active our cells are and whether metabolism gets faster or slower as we grow up and grow old, we need to separate the effects of age and size, which is not easy. You need a large sample with people of all ages, measured with the same methods. Ideally, you'd want measures of total daily energy expenditure, a full tally of the calories used each day.

Researchers have been measuring metabolic rates at rest for more than a century, with some evidence for faster metabolism in children and slower metabolism among the elderly. Yet resting metabolism accounts for only 60 percent or so of the calories we burn over 24 hours and doesn't include the energy we spend on exercise and other physical activity. Online calorie calculators purport to include activity costs, but they're really just a guess based on your self-reported weight and physical activity. In the absence of solid evidence, a kind of folk wisdom has developed, cheered on and cultivated by charismatic hucksters selling metabolic boosters and other snake oil. We're often told our metabolism speeds up at puberty and slows down in middle age, particularly with menopause, and that men have faster metabolisms than women. None of these claims is based on real science.

A METABOLIC DATABASE

MY COLLEAGUES and I have begun to fill that gap in scientific understanding. In 2014 John Speakman, a researcher in metabolism with laboratories at the University of Aberdeen in Scotland and the Chinese Academy of Sciences in Shenzhen, organized an international effort to develop a large metabolic database. Crucially, this database would focus on total daily energy expenditure measured using the doubly labeled water method, an isotope-tracking technique that measures the carbon dioxide produced by the body (and thus the calories burned) over one to two weeks. Doubly labeled water is the gold standard for measuring daily energy expenditures, but it's costly, and you need a specialized lab for the isotope analyses. So although the technique has been around for decades, studies are typically small. Led by Speakman, my lab joined a dozen others around the world in pooling decades of data. We ended up with more than 6,400 measurements of people ranging from babies just eight days old to men and women in their 90s.

Source: "Daily Energy Expenditure through the Human Life Course," by Herman Ponzer et al., in *Science*, Vol. 373; August 2021 (*reference*); Restyled by Jen Christiansen

© 2023 Scientific American

In 2021, after years of collaborative effort, we published the first comprehensive study investigating the effects of age and body size on daily energy expenditure. As expected, we found that metabolic rates increase with body size: bigger people burn more calories. In particular, fat-free mass (the muscles and other organs) is the single strongest predictor of daily energy expenditure. This makes good sense. Fat cells aren't as active as those in the liver, brain, or other tissues, and they don't contribute much to your daily expenditure. More important, with the relation between mass and metabolic rate clearly established from thousands of measurements, we could finally test whether metabolism at each age was faster or slower than we'd expect from size.

The results were a revelation, the first clear road map of metabolism over the human life span.

The results were a revelation, the first clear road map of metabolism over the human life span. We found that, metabolically, babies are born like tiny adults, reflecting their development as part of their mom's energy budget. But metabolism skyrockets over the first year of life, so that by their first birthday children are burning 50 percent more energy than we'd expect for their size. Their cells are far busier than adults' cells, hard at work on growth and development. Earlier studies measuring glucose uptake in the brain during childhood suggest some of this work is neuronal growth and synapse development. Maturation in other systems no doubt contributes as well. Metabolism stays elevated through childhood, slowly decelerating through adolescence to land at adult levels around age 20. Boys decline more slowly than girls, consistent with boys' slower development, but there's no bump at puberty in males or females.

Perhaps the biggest surprise was the stability of our metabolism through middle age. Daily energy expenditures hold remarkably steady from age 20 to 60. No middle age slowdown, no change with menopause. The weight gain so many of us experience in adulthood cannot be blamed on a declining metabolism. As a man in my 40s, I had sort of believed the folk wisdom that metabolism slowed as we aged. My body definitely feels different than it did 10 or 20 years ago. But like hunting some metabolic Sasquatch, when you actually look there's nothing there. Same for the much touted metabolic differences between men and women. Women have lower daily energy expenditures on average, but that is only because women tend to be smaller and carry more of their weight as fat. Compare men and women with the same body weight and body fat percentage, and the metabolic difference disappears.

We did find a decline in metabolism with age, but it doesn't kick in until we hit 60. After 60, metabolism slows by around 7 percent per decade. By the time men and women are in their 90s, their daily expenditures are 20 to 25 percent lower, on average, than those of adults in their 50s. That's after we account for body size and composition. Weight loss with old age, especially diminished muscle mass, compounds the decline in expenditure. As with all age groups, there's a good amount of individual variability. Maintaining a younger, faster metabolism into old age might be a sign of aging well, or perhaps it is even protective against heart disease, dementia and other age-related disease. We can now start to investigate these connections. Guided by our metabolic road map, we have a new world of research ahead of us.

What is already apparent, however, is that a bite of birthday cake does different things for a seven-year-old girl, her middle-aged dad and her elderly grandmother. Clara's bite is likely to be gobbled up by busy cells, fueling development. Mine might go to maintenance, repairing all the little bits of damage accrued through the course of the day. As for Grandma, her aging cells might be slow to use the calories at all, storing them instead as glycogen or fat. Indeed, for any of us, the cake will end up as fat if we eat more calories than we burn.

The road map also highlights a major conundrum of the human condition. Whether they're born into a hunter-gatherer camp, a farming village or an industrial megacity, human youngsters need a lot of help getting food. Other apes learn to forage for themselves by the time they stop nursing, around the age of three or four. Our children are wholly dependent on others for food for years and aren't self-sufficient until their teens. And those least able to fend for themselves have the greatest energy needs. Not only has our species evolved a faster metabolic rate and greater energy demands than other apes, but we must also provision each costly offspring for more than a decade. Where do we get all those calories? Recently my colleagues and I worked out this part of the human energy equation, too.

COSTLY KIDS

THE QUESTION OF CALORIES looms largest in hunter-gatherer and farming communities, where daily life revolves around food production. For most of our species' history, as for most species, there was no other line of work. Every kid knew what they were going to be when they grew up. As late as the mid-1800s, more than half of the American workforce was made up of farmers.

For the past decade I've been working with colleagues to understand the calorie economy in the Hadza community of northern Tanzania. The Hadza are a small population of 1,000 or so, and about half of them maintain a traditional hunting-and-gathering way of life, foraging on the savanna landscape they call home. No population alive today is a perfect model of the past, but groups like the Hadza, who continue these traditions, provide a living example of how these systems work. Men spend most days hunting with bow and arrow or chopping into hollow tree limbs to pillage honey from beehives. Women gather berries and other plant foods or dig for wild tubers in the rocky soil. Hadza camps, small collections of grass houses tucked among the acacia trees, are alive all day with kids being kids, running around, laughing, playing—and waiting for adults to bring them food.

We've measured Hadza energy budgets using doubly labeled water, giving us a clear idea of the calories men and women consume and expend each day. We've also lugged portable respirometry equipment into the bush, a metabolic lab in a briefcase, to measure the energy costs of foraging activities such as walking, climbing, digging tubers and chopping trees. And we've got years of careful observation recording the hours spent each day on different foraging tasks and the amount of food acquired. After more than a decade of work, we've got a complete accounting of the

© 2023 Scientific American

Hadza energy economy: the calories spent to get food, the calories acquired, the proportions shared and consumed.

Tom Kraft of the University of Utah led our team's effort to compare the energy budgets of the Hadza population with similar data from other human groups and from other species of apes. It was a massive project, with researchers poring over old ethnographic accounts of hunter-gatherer and farming groups and combing through ecological studies and doubly labeled water measurements in apes to reconstruct their foraging economies. But when we were finished, what emerged was a new understanding of the energetic foundation for our species' success. We could finally see where all those calories come from, the energy needed to fuel expensive human metabolisms and provision helpless kids.

CLEVER COOPERATORS

IT TURNS OUT humans' unique, cooperative foraging strategy, combined with our clever brains and tools, makes hunting and gathering extremely productive. Even in the harsh, dry savanna of northern Tanzania, Hadza men and women acquire 500 to 1,000 kilocalories of food an hour, on average. Ethnographic records from other groups around the world suggest these rates are typical for hunter-gatherers. Five hours of hunting and gathering can reliably bring in 3,000 to 5,000 kilocalories of food, enough to meet a forager's daily needs and provision the camps' children.

It's the positive feedback engine that propelled the human species to new heights. Hunting and gathering is so productive that it creates an energy surplus. Those extra calories are channeled to offspring, meaning they can take longer to develop, learning skills that make them effective foragers. Reaching adulthood, they'll do just as their parents did, acquiring extra food and plowing those calories into the next generation. Over evolutionary time childhood grows longer as foraging strategies grow more complex. Life spans get extended, too, with natural selection favoring additional years of productive foraging to support children and grandchildren. Grandparents, once rare, become a fixture of the social network.

Apes in the wild are not nearly as productive in gathering food. A forensic accounting of the energy budgets for chimpanzees, gorillas and orangutans shows that males and females get around 200 to 300 kilocalories an hour. It takes them seven hours of foraging just to meet their own needs each day. No wonder they don't share.

Our hyperproductive foraging isn't cheap. People in hunter-gatherer communities expend more than twice as much energy to acquire food as apes in the wild. Surprisingly, human technology and smarts don't make us very energy-efficient. Hadza men and women achieve the same paltry ratio of energy acquired to energy expended that we find in wild apes. Cooperation and culture enable human foragers to be incredibly time-efficient, acquiring lots of calories an hour, but our unique foraging strategies are still energetically demanding. Hunting and gathering is hard work.

Farming isn't any easier, but our analyses found it can be even more productive. When we compared the energy budgets for the Hadza and other hunter-gatherer populations with those of traditional farming groups, we found that farmers typically produce far more calories an hour. The Tsimane community, a population in the Amazonian rain forest of Bolivia, provides a useful point of comparison. The Tsimane get most of their calories from farming, but they also hunt, fish and collect wild plants. With farmed foods as their energy staple, they produce nearly twice as many calories an hour as the Hadza. They're more energy-efficient as well, getting more food from every calorie they spend foraging and farming.

Those extra calories are embodied in the children running around Tsimane villages. More food and faster production mean a lighter workload for mothers because others in the community can more easily share the time and energy costs of caring for kids. As with many subsistence farming communities, Tsimane families tend to be large. Women have an average of nine children over the course of their lives. Compare that with the average fertility rate of six children per mother in the Hadza community, and the impact of that extra energy is inescapable. And it's not just the Tsimane. Farming communities tend to have higher fertility rates than hunter-gatherer communities. Increased fertility is an important reason farming overtook hunting and gathering in the Neolithic age, the time spanning roughly 12,000 to 6,500 years ago. Archaeological sites across Eurasia and the Americas document a rising tide of children and adolescents following the development of agriculture.

HAVING OUR CAKE

FROM THIS PERSPECTIVE, a kid's birthday party is more than a personal milestone. It's a celebration of our improbable evolutionary story. There's the food, of course. We get the flour and sugar for the cake from our farming ancestors, the fire to bake it from the Paleolithic era. The milk and eggs come from animals that we've completely transformed from species we once hunted, shaped to our will over generations of careful husbandry. And there's the calendar we use to mark our days and measure our years, an invention of agriculturalists who needed to know precisely when to reap and sow. Hunter-gatherers track the seasons and lunar cycles but have little use for accurate annual calendars. There are no birthdays in a Hadza camp.

But the key element of any celebration is the community of friends and relatives, multiple generations gathering to eat and laugh and sing. Our evolved social contract—to hunt, gather and farm collectively—tied us together, gave us our childhood and extended our golden years. Cooperative foraging also helped to fuel the cultural complexity and innovation that make birthdays and other rituals so fantastical and diverse. And at the center of it all is the universal commitment to share.

With eight billion humans on the planet today, one might begin to worry that we've taken things a bit too far. We've learned to turbocharge our energy budgets by tapping into climate-changing fossil fuels and flooding our world with cheap food. Calories are so easy to produce that very few of us spend our days foraging, a first in the history of life. This massive shift has been a boon to our collective creativity, enabling many to spend their lives as artists, doctors, teachers, scientists—a range of careers outside of food production. Having carved out our own strange niche, far removed from the laws that govern the rest of the natural world, we have only ourselves to look to for guidance. With a little luck and a lot of cooperation, we just might secure the human lineage another couple million birthdays. Make a wish. SA

Herman Pontzer is a professor of evolutionary anthropology at Duke University. He studies how evolution has shaped human physiology and health. He is author of *Burn: New Research Blows the Lid Off How We Really Burn Calories, Stay Healthy, and Lose Weight* (Avery, 2021).

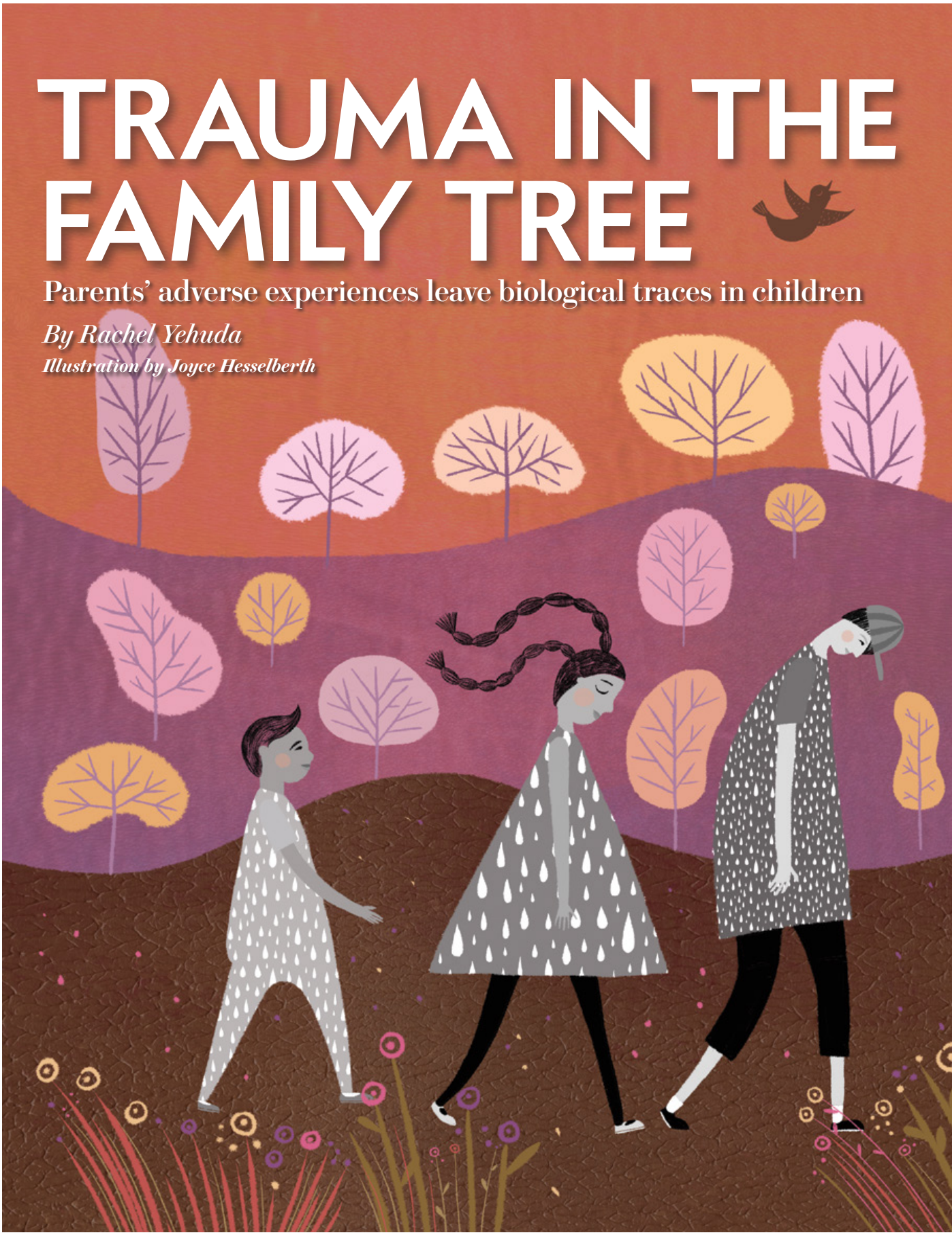
© 2023 Scientific American

TRAUMA IN THE FAMILY TREE

Parents' adverse experiences leave biological traces in children

By Rachel Yehuda

Illustration by Joyce Hesselberth



© 2023 Scientific American

© 2023 Scientific American

AFTER THE TWIN TOWERS OF THE WORLD TRADE CENTER collapsed on September 11, 2001, in a haze of horror and smoke, clinicians at the Icahn School of Medicine at Mount Sinai in Manhattan offered to check anyone who'd been in the area for exposure to toxins. Among those who came in for evaluation were 187 pregnant women. Many were in shock, and a colleague asked if I could help diagnose and monitor them. They were at risk of developing posttraumatic stress disorder, or PTSD—experiencing flashbacks, nightmares, emotional numbness or other psychiatric symptoms for years afterward. And were the fetuses at risk?

My trauma research team quickly trained health professionals to evaluate and, if needed, treat the women. We monitored them through their pregnancies and beyond. When the babies were born, they were smaller than usual—the first sign that the trauma of the World Trade Center attack had reached the womb. Nine months later we examined 38 women and their infants when they came in for a wellness visit. Psychological evaluations revealed that many of the mothers had developed PTSD. And those with PTSD had unusually low levels of the stress-related hormone cortisol, a feature that researchers were coming to associate with the disorder.

Surprisingly and disturbingly, the saliva of the nine-month-old babies of the women with PTSD also showed low cortisol. The effect was most prominent in babies whose mothers had been in their third trimester on that fateful day. Just a year earlier a team I led had reported low cortisol levels in adult children of Holocaust survivors, but we'd assumed that it had something to do with being raised by parents who were suffering from the long-term emotional consequences of severe trauma. Now it looked like trauma could leave a trace in offspring even before they are born.

In the decades since, research by my group and others has confirmed that adverse experiences may influence the next generation through multiple pathways. The most apparent route runs through parental behavior, but influences during gestation and even changes in eggs and sperm may also play a role. And all these channels seem to involve epigenetics: alterations in the way that genes function. Epigenetics potentially explains why effects of trauma may endure long after the immediate threat is gone, and it is also implicated in the diverse pathways by which trauma is transmitted to future generations.

The implications of these findings may seem dire, suggesting that parental trauma predisposes offspring to be vulnerable to mental health conditions. But there is some evidence that the epigenetic response may serve as an adaptation that might help the children of traumatized parents cope with similar adversities. Or could both outcomes be possible?

IN THE AFTERMATH

MY FIRST ENCOUNTER with intergenerational transmission of trauma was in the 1990s, soon after my team documented high rates of PTSD among Holocaust survivors in my childhood community in Cleveland. The first study of its kind, it garnered a lot of publicity; within weeks I found myself heading a newly created Holocaust research center at Mount Sinai staffed largely by professional volunteers. The phone was ringing off the hook. The callers weren't all Holocaust survivors, though; most were the adult children of Holocaust survivors. One particularly persistent caller—I'll call him Joseph—insisted that I study people like him. "I'm a casualty of the Holocaust," he claimed.

When he came in for an interview, Joseph didn't look like a casualty of anything. A handsome and wealthy investment banker in an Armani suit, he could've stepped off the pages of a magazine. But Joseph lived each day with a vague sense that something terrible was going to happen and that he might need to flee or fight for his life. He'd been preparing for the worst

© 2023 Scientific American

since his early 20s, keeping cash and jewelry at hand and becoming proficient in boxing and martial arts. Lately he was tormented by panic attacks and nightmares of persecution, possibly triggered by reports of ethnic cleansing in Bosnia.

Joseph's parents had met in a displaced-persons camp after surviving several years at Auschwitz, then arrived penniless in the U.S. His father worked 14 hours a day and said very little, never mentioning the war. But almost every night he woke the family with shrieks of terror from his nightmares. His mother spoke endlessly about the war, telling vivid bedtime stories about how relatives had been murdered before her eyes. She was determined that her son succeed, and his decision to remain unattached and childless infuriated her. "I didn't survive Auschwitz so that my own child would end the family line," she'd say. "You have an obligation to me and to history."

We ended up talking to many people like Joseph: adult children of Holocaust survivors who suffered from anxiety, grief, guilt, dysfunctional relationships and intrusions of Holocaust-related imagery. Joseph was right—I needed to study people like him. Because those who were calling us were (in research-speak) self-selecting, we decided to evaluate the offspring of the Holocaust survivors we had just studied in Cleveland. The results were clear. Survivors' adult children were more likely than others to have mood and anxiety disorders, as well as PTSD. Further, many Holocaust offspring also had low cortisol levels—something that we had observed in their parents with PTSD.

FIGHT, FLIGHT—OR FREEZE

WHAT DID IT ALL MEAN? Unraveling the tangle of trauma, cortisol and PTSD has occupied me and many other researchers for the decades since. In the classic fight-or-flight response, identified in the 1920s, a threatening encounter triggers the release of stress hormones such as adrenaline and cortisol. The hormones prompt a cascade of changes, such as quickening the pulse and sharpening the senses to enable the threatened person or animal to focus on and react to the immediate danger. These acute effects were believed to dissipate once the danger receded.

In 1980, however, psychiatrists and other advocates for Vietnam War veterans won a prolonged struggle to get post-traumatic stress included in the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-III*). It was the first official recognition that trauma could have long-lasting effects. But the diagnosis was controversial. Many psychologists believed that its inclusion in *DSM-III* had been politically, rather than scientifically, driven—in part because there were no scientific explanations for how a threat could continue to influence the body long after it was removed.

Complicating matters, studies of Vietnam veterans were generating perplexing results. In the mid-1980s neuroscientists John Mason, Earl Giller and Thomas Kosten, then all at Yale University, reported that veterans with PTSD had high levels of adrenaline but lower levels of cortisol than patients with other psychiatric diagnoses. Because stress usually causes stress hormones, including cortisol, to rise, many researchers, including myself, were skeptical of these observations. When I joined the Yale laboratory as a postdoctoral fellow a year later, I studied a different group of veterans using other methods for measuring cortisol. To my astonishment, I replicated the finding.

I still couldn't believe that the low cortisol levels had anything to do with trauma. Surely the Holocaust was as terrible as the Vietnam War, I reasoned—and growing up as a rabbi's daughter in a community full of Holocaust survivors, many of them my friends' parents, I'd noticed nothing out of the ordinary about them. I was sure that they didn't have either PTSD or low cortisol, I told my mentor, Giller. "That's a testable hypothesis," he responded. "Why don't you study that, instead of conjecturing?"

So my team of five people landed in Cleveland, along with a centrifuge and other equipment. We stayed in my parents' home, walking door to door to interview people by day and returning to test blood and urine samples in the evening. When the results came in, they were clear: half the Holocaust survivors had PTSD, and those with PTSD had low cortisol. There was no question about it—even if the traumatic experience happened long ago, PTSD went hand in hand with low cortisol.

> A shape was emerging that connected childhood adversity with low cortisol and the possibility of future PTSD.

But why? And which came first? An important clue came from a 1984 review by the late Allan Munck and other researchers at the Geisel School of Medicine at Dartmouth. They noted that among stress hormones, cortisol played a special, regulatory role. High levels of stress hormones, if sustained for a long time, harm the body in multiple ways, weakening the immune system and increasing susceptibility to problems such as hypertension. But in a context of acute trauma, cortisol may paradoxically also have a protective effect. It shuts down the release of stress hormones—including itself—reducing the potential damage to organs and the brain. Such a trauma-induced feedback loop could conceivably reset the cortisol "thermostat" to a lower level.

I picked up another piece of the puzzle and placed it. In the early 1990s we'd shown that Vietnam veterans were more likely to develop PTSD if they'd been abused as children. Slowly a shape was emerging that connected intense childhood adversity—a period of "freeze" because a child usually cannot fight or flee—with low cortisol and the possibility of future PTSD. We studied people who'd been raped or who'd been in auto accidents when they came into emergency rooms, finding that those with lower cortisol levels were more likely to develop PTSD after the attack or accident.

Could low cortisol levels have been present before the event that brought them into the emergency room? I wondered. If someone with low cortisol was subjected to a traumatic experience, we reasoned, the cortisol levels in their bodies might be too low to tamp down the stress reaction. Adrenaline levels might then shoot way up, searing the memory of the new trauma into the brain—from where it might later surface as flashbacks or nightmares. Perhaps low cortisol marked a vulnerability to developing PTSD.

The study of Holocaust offspring supported this conjecture. Children of Holocaust survivors with PTSD tended to have low

© 2023 Scientific American

cortisol even if they did not have their own PTSD. As we'd suspected, low cortisol seemed related to vulnerability to PTSD.

THE FEEDBACK LOOP

BUT WHAT MECHANISM connected trauma exposure to low cortisol to future PTSD? We began a series of studies to answer this question. Significantly, we found that Vietnam veterans with PTSD had a greater number of glucocorticoid receptors. These are proteins to which cortisol binds to exert its diverse influences. That suggested a greater sensitivity to cortisol: a small increase in the hormone's concentration would precipitate a disproportionate physiological reaction. But it wasn't until we looked more closely at the molecular underpinnings of cortisol functioning—in part by examining epigenetics—that we understood how exposure to trauma might reset the cortisol feedback loop.

In the 1990s scientists were realizing that the output of our genes is sensitive to factors not written directly into our genetic code. Genes provide the templates for producing proteins. But much like cakes baked using the same ingredients may turn out differently depending on variations in the oven's temperature, how much of those proteins gets produced, or "expressed," depends on the environment. The discovery gave rise to epigenetics, the study of what influences gene expression and how. It proved crucial to understanding both the neurobiology of PTSD and the intergenerational effects of trauma.

Epigeneticists explore the switches that turn gene expression on and off. One such mechanism, called methylation, involves a methyl group—a methane molecule that is missing one of its four hydrogen atoms, leaving a chemical bond free to attach to another atom or molecule. Methylation is a process by which, in the presence of specific enzymes, methyl groups attach to key sites on a strand of DNA or within the complex of DNA and proteins known as chromatin. By occupying these sites like roadblocks on a highway, methyl groups can alter transcription, a basic step in gene expression where a piece of RNA is made from a DNA template. Increased methylation generally impedes RNA transcription, whereas less methylation enhances transcription. These changes are enduring in that they survive normal cell division and require specific enzymes for their removal.

In 2015 our group became one of the first to pinpoint epigenetic changes on stress-related genes of veterans with PTSD. These alterations partially explained why trauma's effects were so persistent, lasting for decades. Specifically we observed reduced methylation in an important region of *NR3C1,* a gene that encodes the glucocorticoid receptor, likely increasing the sensitivity of these receptors.

This epigenetic modification suggests a potential explanation for how trauma might reset cortisol levels. The body regulates the stress response through a complicated feedback mechanism. A rise in cortisol levels will prompt the body to produce less of the hormone, which may drive up the numbers and responsiveness of glucocorticoid receptors. Given the epigenetic and other changes occurring with sustained responses to trauma, the feedback loop might become recalibrated. In people who have already endured trauma, their stress systems might be sensitized and their cortisol levels diminished—increasing their adrenaline response to further trauma and leading to PTSD.

EPIGENETIC INHERITANCE

COULD SOME OF THESE epigenetic changes in trauma survivors also be found in the children of trauma survivors? Finding low cortisol in the 9/11 babies back in 2002 had told us that we'd been thinking about some things all wrong. We'd assumed all along that trauma was behaviorally transmitted: Joseph's problems seemed to result from the stressful, bereaved atmosphere in his childhood home. But now it looked like the uterine environment also played a role. So did the sex of the traumatized parent.

In our early studies of Holocaust offspring, we had selected only those people with two parents who were Holocaust survivors. We redid the studies to figure out if the sex of the parent mattered—and it did. Those whose mother (or both parents) had PTSD tended to exhibit lower cortisol levels and showed evidence of more sensitive glucocorticoid receptors. In contrast, those whose fathers, but not mothers, had PTSD showed the opposite effect.

Taking a closer look, we again discovered lower methylation within the glucocorticoid receptor gene, *NR3C1,* in Holocaust offspring whose mothers, or both parents, had PTSD. These changes mirrored what we'd observed in the maternal survivors themselves. But in offspring with only paternal PTSD, we observed more methylation—the opposite effect. These findings raised the possibility that PTSD in mothers and fathers might lead to different epigenetic changes on the glucocorticoid receptor in children.

In a second series of studies beginning in 2016, we examined methylation within another gene, *FKBP5,* which encodes a protein involved in regulating the ability of the glucocorticoid receptor to bind cortisol. The findings showed related methylation patterns within the *FKBP5* gene in both Holocaust parents and their children. But because of the small number of participants in that study—by this time it was difficult to find many living Holocaust survivors who could participate with their offspring—we couldn't examine how factors such as the parents' PTSD status might contribute to *FKBP5* methylation.

We were able to replicate and extend this work in a substantially larger sample of just the Holocaust offspring, however. In 2020 we reported lower levels of *FKBP5* methylation in the adult children whose mothers—and not fathers—were exposed to the Holocaust during childhood. This effect was independent of whether the mother had PTSD. It suggested that trauma might have affected the mothers' eggs decades before her children were conceived, while she was herself a child.

Given the obvious difficulties in studying generations of people, scientists often resort to animal studies to explore epigenetic transmission. In 2014 Brian Dias and Kerry Ressler, then both at the Emory University School of Medicine, reported an intergenerational epigenetic pathway that ran through sperm. They gave a male mouse a mild electric shock as it smelled a cherry blossom scent, stimulating a fear response to the odor. The response was accompanied by epigenetic changes in its brain and sperm. Intriguingly, the male offspring of the shocked mice demonstrated a similar fear of cherry blossoms—as well as epigenetic changes in their brain and sperm—without being exposed to the shock. These effects were passed down for two generations. In other words, the lesson the grandfather mouse learned, that the cherry blossom scent means danger, was transmitted to its son and grandson.

In a study published in 2020, my colleagues and I experi-

© 2023 Scientific American

mented with genome-wide gene expression, a tool that can identify links between protein expression and specific conditions across the entire human genome. With this approach, we again observed distinct patterns of gene expression linked with maternal and paternal trauma exposure and PTSD.

IN THE WOMB

APART FROM ALTERING the eggs and sperm that encapsulate our genetic inheritance, sometimes decades before conception, trauma also seems to influence the uterine environment. Meticulous studies of the offspring of women who were pregnant during the Dutch Famine—a six-month period during World War II when the Nazis blocked the food supply to the Netherlands, causing widespread starvation—provided an early indication of in utero effects. Researchers discovered that the combined effects of extreme stress and nutritional deprivation, such as deficits in metabolism and susceptibility to cardiovascular illness, depended on the trimester of exposure.

The 9/11 babies were also impacted in the womb, with those in the third trimester having significantly lower cortisol levels. What this condition meant for their future development I sadly never found out. At the wellness visit, mothers who had PTSD (and low cortisol) were more likely to report that their nine-month-olds were unusually anxious and afraid of strangers. But we didn't get the funding to follow the babies into adulthood.

How might the uterine environment leave a trauma trace in the offspring? Our work on Holocaust survivors and their adult children provided some clues. The story is again complicated, and it involves an enzyme known as 11-beta-hydroxysteroid dehydrogenase type 2 (11β-HSD2). Holocaust survivors had lower levels of the enzyme than those who hadn't lived through the Holocaust—and such effects were particularly pronounced in those who were the youngest during World War II. The enzyme is normally concentrated in the liver, kidneys and brain. Under conditions of food deprivation, however, the body can lower levels of 11β-HSD2 to increase metabolic fuel in the interest of promoting survival. In adults, the enzyme level will return to what it had been originally when there is no more starvation, but in children, the level may remain low. Our findings suggested that 11β-HSD2 levels might have been altered during childhood when Holocaust survivors were exposed to long periods of malnourishment; the change persisted well into old age.

In the children of women who were Holocaust survivors, however, we saw quite the opposite: 11β-HSD2 levels were higher than in Jewish control subjects. The result might seem contradictory, but there is a logic to it. During pregnancy, 11β-HSD2 also acts in the placenta, protecting the fetus from exposure to circulating maternal cortisol, which can be toxic to the developing brain. The enzyme, which is particularly active in the third trimester, converts maternal cortisol into an inactive form, creating a kind of chemical shield in the placenta that protects the fetus from the hormone's harmful effects. The high levels of this enzyme in the offspring of Holocaust survivors may thus reflect an adaptation, an effort to protect the fetus from the lowered 11β-HSD2 levels in their mothers.

All of this means that offspring are not always passive recipients of their parents' scars. Just as a parent was able to survive trauma by means of biological adaptations, offspring can sometimes adapt to the biological impact of their parents' trauma.

How traumatized parents interact with their children, of course, also influences their development. One of the most powerful nonfiction accounts of growing up with Holocaust survivor parents was Art Spiegelman's graphic novel *Maus;* serialized between 1980 and 1991, it broke through a cultural barrier, helping others to open up about their suffering. Many psychologists and neuroscientists have examined the traumatized family, finding ever more subtleties, and the story will continue to unfold for decades to come.

An important question is whether epigenetic alterations in stress-related genes, particularly those reflected in the offspring of traumatized parents, are necessarily markers of vulnerability or whether they may reflect a mechanism through which offspring become better equipped to cope with adversity. This is an area we're actively exploring.

Epigenetic influences might represent the body's attempts to prepare offspring for challenges similar to those encountered by their parents.

It is tempting to interpret epigenetic inheritance as a story of how trauma results in permanent damage. Epigenetic influences might nonetheless represent the body's attempts to prepare offspring for challenges similar to those encountered by their parents. As circumstances change, however, the benefits conferred by such alterations may wane or even result in the emergence of novel vulnerabilities. Thus, the survival advantage of this form of intergenerational transmission depends in large part on the environment encountered by the offspring themselves.

Moreover, some of these stress-related and intergenerational changes may be reversible. Several years ago we discovered that combat veterans with PTSD who benefited from cognitive-behavioral psychotherapy showed treatment-induced changes in *FKBP5* methylation. The finding confirmed that healing is also reflected in epigenetic change. And Dias and Ressler reconditioned their mice to lose their fear of cherry blossoms; the offspring conceived after this "treatment" did not have the cherry blossom epigenetic alteration, nor did they fear the scent. Preliminary as they are, such findings represent an important frontier in psychiatry and may suggest new avenues for treatment.

The hope is that as we learn more about the ways catastrophic experiences have shaped both those who lived through those horrors and their descendants, we will become better equipped to deal with dangers now and in the future, facing them with resolution and resilience. SA

Rachel Yehuda is a professor of psychiatry and neuroscience and director of the Center for Psychedelic Psychotherapy and Trauma Research at the Icahn School of Medicine at Mount Sinai. She is also director of mental health at the James J. Peters Veterans Affairs Medical Center.

© 2023 Scientific American

Messenger RNA Therapies Finally Arrive

Instructing our cells to make specific proteins could control influenza, autoimmune diseases, even cancer

By Drew Weissman

IN JUST 17 YEARS, MESSENGER RNA therapies went from proof of concept to global salvation. The Pfizer-BioNTech and Moderna vaccines for COVID-19 have been given to hundreds of millions of people, saving countless lives.

In 2005 Katalin Karikó and I created a way to make mRNA molecules that would not cause dangerous inflammation when injected into an animal's tissue. In 2017 Norbert Pardi and I demonstrated that modified mRNA, carried into human cells by a fatlike nanoparticle, was protected from being broken down by the body and prompted immune cells to generate antibodies that neutralize an invading virus more effectively than the immune system could on its own. The COVID vaccines made by Pfizer-BioNTech and Moderna both use this mRNA-liquid-nanoparticle "platform"—known as mRNA-LNP. In large clinical trials, it prevented more than 90 percent of the people who received the vaccines from becoming ill.

These extremely promising trials, and massive studies of people who have since received the vaccines, have finally given us sufficient information about the safety and efficacy of mRNA vaccines in humans. The platform outperformed more conventional approaches, in which vaccines are grown in laboratory cell cultures or chicken eggs. The rapid development also accelerated investment in further research. And because the U.S. Food and Drug Administration and similar regulatory agencies are now familiar with the technology, assessment of new therapeutics should come readily.

Messenger RNA vaccines instruct cells to create proteins that induce an immune response to an invader such as the SARS-CoV-2 virus, training the immune system to attack the actual pathogen if it infects the person in the future. These vaccines are easier to produce in large quantities than conventional protein therapies (genetically engineered versions of natural human or pathogen proteins) and monoclonal antibody therapies (lab-produced molecules that attack viruses in the same way that human antibodies do). And once a reliable manufacturing facility is built, it can quickly switch to a new mRNA vaccine or drug—unlike protein or monoclonal facilities, which must reengineer production from the ground up for each new therapy.

Success has inspired researchers, companies and government labs to pursue mRNA therapies for many infectious diseases, including influenza, cytomegalovirus, herpes simplex virus 2, norovirus, rabies, malaria, tuberculosis, dengue, Zika, HIV, hepatitis C and the entire family of coronaviruses. In each case, researchers wanted to determine exactly how mRNA-LNP vaccines induce potent antibody responses.

Work on mRNA vaccines also has expanded to certain cancers, food and environmental allergies, and autoimmune diseases. Positive results against ATTR amyloidosis, a fatal condition that involves the liver, have already been produced in a phase 1 clinical trial. Although the use of protein-based medications for certain illnesses is expanding quickly, large doses are typically required, and production is often difficult and expensive; mRNA delivery of therapeutic proteins could help. The approach has already worked in animals for issues as disparate as bone repair and asthma, and some treatments have advanced to human clinical trials. The Defense Advanced Research Projects Agency has even experimented with mRNA delivery of monoclonal antibodies that could be tailored for previously unidentified infectious diseases, with the goal of supplying reliable manufacturing of such remedies within 60 days.

The concentrated COVID work has also helped make mRNA a leader in nucleic acid therapeutics—approaches that can produce nearly any protein made by a specific cell. The technique is starting to be applied, and it could fight diseases in more convenient, less invasive and less expensive ways. For example, the FDA has

© 2023 Scientific American

approved gene therapy for sickle cell anemia, and it is working in the U.S., although it requires marrow to be extracted from a person's bone, treated and reinserted; mRNA therapy could be delivered to marrow with a straightforward injection into a person's arm. If that works, sickle cell treatment could be greatly expanded in countries where the condition is widespread.

In a similar fashion, mRNA therapeutics could revolutionize the treatment of many infectious diseases in developing countries, greatly improving health-care equity. I have collaborated with labs around the world. Thai investigators at the vaccine center at Chulalongkorn in Bangkok and I made a Thai COVID vaccine and established a quality manufacturing center to produce it for Thailand and seven surrounding low- and middle-income countries. I undertook similar work in Africa and eastern Europe, with plans to extend it to South America.

Plenty of hurdles remain, including the creation of a better supply chain for delivering raw mRNA vaccine and materials needed for its production worldwide, as well as improvements that could reduce the dosage a person needs to receive. Yet the ease of mRNA production should enable most countries to make their own medications, as long as they can attract and retain researchers who can develop subsequent therapeutics that in turn keep domestic, high-quality manufacturing sites operating.

Drew Weissman is a professor of vaccine research at the University of Pennsylvania. The nucleoside-modified mRNA-lipid-nanoparticle vaccine platform his laboratory created is used in COVID-19 vaccines made by Pfizer-BioNTech and Moderna. Weissman receives royalties from a patent for nucleoside-modified mRNA that is licensed by those two companies.

Illustration by Thomas Fuchs

© 2023 Scientific American

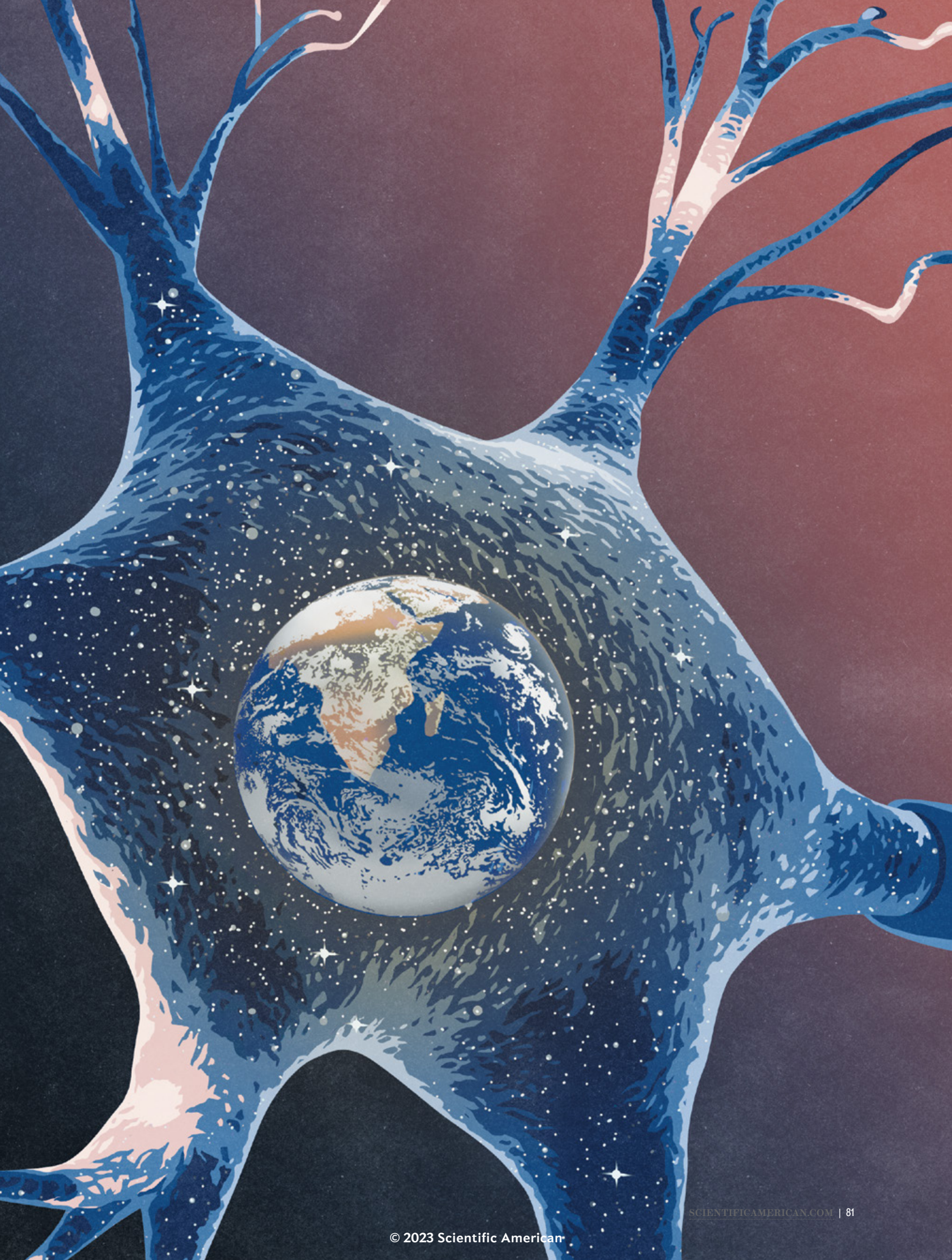
Constructing the World from Inside Out

The brain probes your physical surroundings to select just the information needed to survive and flourish

By György Buzsáki

Illustration by Stefania Infante

© 2023 Scientific American



© 2023 Scientific American

AS A YOUNG COURSE INSTRUCTOR IN SEMINARS for medical students, I faithfully taught neurophysiology by the book, enthusiastically explaining how the brain perceives the world and controls the body. Sensory stimuli from the eyes, ears, and such are converted to electrical signals and then transmitted to the relevant parts of the sensory cortex that process these inputs and induce perception. To initiate a movement, impulses from the motor cortex instruct the spinal cord neurons to produce muscular contraction.

Most students were happy with my textbook explanations of the brain's input-output mechanisms. Yet a minority—the clever ones—always asked a series of awkward questions. "Where in the brain does perception occur?" "What initiates a finger movement before cells in the motor cortex fire?" I would always dispatch their queries with a simple answer: "That all happens in the neocortex." Then I would skillfully change the subject or use a few obscure Latin terms that my students did not really understand but that seemed scientific enough so that my authoritative-sounding accounts temporarily satisfied them.

Like other young researchers, I began my investigation of the brain without worrying much whether this perception-action theoretical framework was right or wrong. I was happy for many years with my own progress and the spectacular discoveries that gradually evolved into what became known in the 1960s as the field of "neuroscience." Yet my inability to give satisfactory answers to the legitimate questions of my smartest students has haunted me ever since. I had to wrestle with the difficulty of trying to explain something that I didn't really understand.

Over the years I realized that this frustration was not uniquely my own. Many of my colleagues, whether they admitted it or not, felt the same way. There was a bright side, though, because these frustrations energized my career. They nudged me over the years to develop a perspective that provides an alternative description of how the brain interacts with the outside world.

The challenge for me and other neuroscientists involves the weighty question of what, exactly, is the mind. Ever since the time of Aristotle, thinkers have assumed that the soul or the mind is initially a blank slate, a tabula rasa on which experiences are painted. This view has influenced thinking in Christian and Persian philosophies, British empiricism and Marxist doctrine. In the past century it has also permeated psychology and cognitive science. This "outside-in" view portrays the mind as a tool for learning about the true nature of the world. The alternative view—one that has defined my research—asserts that the primary preoccupation of brain networks is to maintain their own internal dynamics and perpetually generate myriad nonsensical patterns of neural activity. When a seemingly random action offers a benefit to an organism's survival, the neuronal pattern leading to that action gains meaning. When an infant utters "te-te," the parent happily offers the baby "teddy," so the sound "te-te" acquires the meaning of the teddy bear. Recent progress in neuroscience has lent support to this framework.

DOES THE BRAIN "REPRESENT" THE WORLD?

NEUROSCIENCE INHERITED the blank slate framework millennia after early thinkers gave names like tabula rasa to mental operations. Even today we still search for neural mechanisms that might relate to their dreamed-up ideas. The dominance of the outside-in framework is illustrated by the outstanding discoveries of the legendary scientific duo David Hubel and Torsten Wiesel, who introduced single-neuronal recordings to study the visual system and were awarded the Nobel Prize in Physiology or Medicine in 1981. In their signature experiments, they recorded neural activity in animals while showing them images of various shapes. Moving lines, edges, light or dark areas,

© 2023 Scientific American

and other physical qualities elicited firing in different sets of neurons. The assumption was that neuronal computation starts with simple patterns that are synthesized into more complex ones. These features are then bound together somewhere in the brain to represent an object. No active participation is needed. The brain automatically performs this exercise.

The outside-in framework presumes that the brain's fundamental function is to perceive "signals" from the world and correctly interpret them. But if this assumption is true, an additional operation is needed to respond to these signals. Wedged between perceptual inputs and outputs resides a hypothetical central processor—which takes in sensory representations from the environment and makes decisions about what to do with them to perform the correct action.

So what exactly is the central processor in this outside-in paradigm? This poorly understood and speculative entity goes by various names—free will, homunculus, decision maker, executive function, intervening variables or simply just a "black box." It all depends on the experimenter's philosophical inclination and whether the mental operation in question is applied to the human brain, brains of other animals, or computer models. Yet all these concepts refer to the same thing.

An implicit practical implication of the outside-in framework is that the next frontier for progress in contemporary neuroscience should be to find where the putative central processor resides in the brain and systematically elaborate the neuronal mechanisms of decision-making. Indeed, the physiology of decision-making has become one of the most popular focuses in contemporary neuroscience. Higher-order brain regions, such as the prefrontal cortex, have been postulated as the place where "all things come together" and "all outputs are initiated." When we look more closely, however, the outside-in framework does not hold together.

This approach cannot explain how photons falling on the retina are transformed into a recollection of a summer outing. The outside-in framework requires the artificial insertion of a human experimenter who observes this event [*see box at right*]. The experimenter-in-the-middle is needed because even if neurons change their firing patterns when receptors on sensory organs are stimulated—by light or sound, for instance—these changes do not intrinsically "represent" anything that can be absorbed and integrated by the brain. The neurons in the visual cortex that respond to the image of, say, a rose have no clue. They do not "see" the appearance of a flower. They simply generate electrical oscillations in response to inputs from other parts of the brain, including those arriving along multiple complex pathways from the retina.

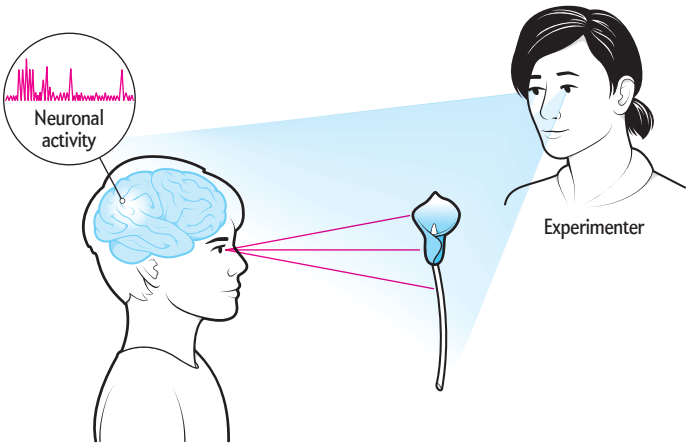
In other words, neurons in sensory cortical areas and even in the hypothetical central processor cannot "see" events that happen in the world. There is no interpreter in the brain to assign meaning to these changes in neuronal firing patterns. Without a magi-

Outside In vs. Inside Out

The idea of the brain as a blank slate onto which experience is written has existed since ancient times—and persists today in modified form. Some neuroscientists have begun to question this theory because it requires a hard-to-justify assumption about the way we perceive and process events from the outside world—in particular, the need to involve a hypothetical "interpreter" to explain what is happening.

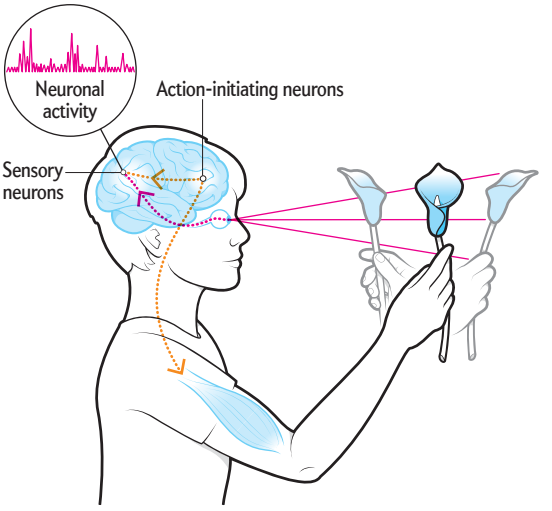
Outside-In Framework

A stimulus—the image of a flower—reaches the eyes, and the brain responds by causing neurons to fire. This theory is plausible only with the involvement of an "experimenter" to observe and establish a relation between the flower and the neuronal responses it induces. Absent the experimenter, neurons in the sensory cortex do not "see" the flower.



Inside-Out Framework

The alternative, inside-out theory does away with the experimenter. It presumes instead that we come to understand the external world by taking actions—moving a flower, for instance—to learn about an object. To accomplish this task, inputs from action-initiating neurons combine with sensory inputs to provide an understanding of the object's size, shape and other attributes. A meaningful picture arises, allowing the neurons to "see" the flower.



Graphic by Brown Bird Design

© 2023 Scientific American

cal homunculus watching the activities of all the neurons in the brain with the omniscience of the experimenter, the neurons that take this all in are unaware of the events that caused these changes in their firing patterns. Fluctuations in neuronal activity are meaningful only for the scientist who is in the privileged position of observing both events in the brain and events in the outside world and then comparing the two perspectives.

PERCEPTION IS WHAT WE DO

BECAUSE NEURONS have no direct access to the outside world, they need a way to compare or "ground" their firing patterns to something else. The term "grounding" refers to the ability of the brain's circuits to assign meaning to changes in neuronal firing patterns that result from sensory inputs. They accomplish this task by relating this activity to something else. The "dah-dah-dit" Morse code pattern becomes meaningful only when it has previously been linked to the letter "G." In the brain, the only available source of a second opinion appears when we initiate some action.

We learn that sticks that look bent in water are not broken by moving them. Similarly, the distances between two trees and two mountain peaks may appear identical, but by moving around and shifting our perspective we learn the difference.

The outside-in framework follows a chain of events from perception to decision to action. In this model, neurons in dedicated sensory areas are "driven" by environmental signals and thus cannot relate their activity to something else. But the brain is not a serial processing unit; it does not proceed one by one through each of these steps. Instead any action a person takes involves the brain's motor areas informing the rest of the cerebral cortex about the action initiated—a message known as a corollary discharge.

Neuronal circuits that initiate an action dedicate themselves to two tasks. The first is to send a command to the muscles that control the eyes and other bodily sensors (the fingers and tongue, among others). These circuits orient bodily sensors in the optimal direction for in-depth investigation of the source of an input and enhance the brain's ability to identify the nature and location of initially ambiguous incoming signals from the senses.

The second task of these same action circuits involves sending notifications—the corollary discharges—to sensory and higher-order brain areas. Think of them as registered mail receipts. Neurons that initiate eye movement also notify visual sensory areas of the cortex about what is happening and disambiguate whether, say, a flower is moving in the wind or being handled by the person observing it.

This corollary message provides the second opinion sensory circuits need for grounding—a confirmation that "my own action is the agent of change." Similar corollary messages are sent to the rest of the brain when a person takes actions to investigate the flower and its relationship to that person and other objects. Without such exploration, stimuli from the flower alone—the photons arriving on the retina connected to an inexperienced brain—would never become signals that furnish a meaningful description of the flower's size and shape. Perception then can be defined as what we *do*—not what we passively take in through our senses.

You can demonstrate a simple version of the corollary discharge mechanism. Cover one of your eyes with one hand and press against the side of the other eye gently with the tip of your finger at about three times per second while reading this text. You will see immediately that the page is moving back and forth. By comparison, when you move your eyes naturally to read or look around the room, nothing else seems to move. This constancy occurs because neurons that initiate eye movements to scan sentences also send a corollary signal to the visual system to indicate whether the world or the eyeball is moving, thus stabilizing the perception of your surroundings.

LEARNING BY MATCHING

THE CONTRAST between outside-in and inside-out approaches becomes most striking when used to explain the mechanisms of learning. A tacit assumption of the blank slate model is that the complexity of the brain grows with the amount of experience. As we learn, the interactions of brain circuits should become increasingly more elaborate. In the inside-out framework, however, experience is not the main source of the brain's complexity.

Instead the brain organizes itself into a vast repertoire of preformed patterns of firing known as neuronal trajectories. This self-organized brain model can be likened to a dictionary filled initially with nonsensical words. New experience does not change the way these networks function—their overall activity level, for instance. Learning takes place, rather, through a process of matching the preexisting neuronal trajectories to events in the world.

To understand the matching process, we need to examine the advantages and constraints brain dynamics impose on experience. In the basic version, models of blank slate neuronal networks assume a collection of largely similar, randomly connected neurons. The presumption is that brain circuits are highly plastic and that any arbitrary input can alter the activity of neuronal circuits.

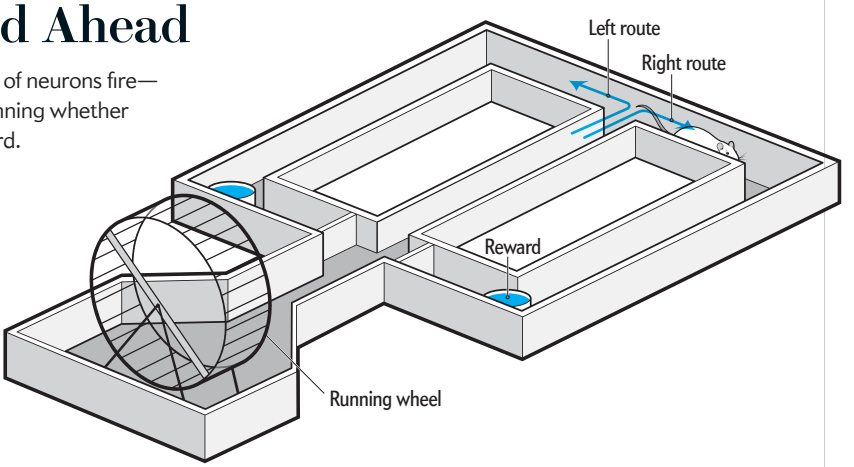
We can see the fallacy of this approach by considering an example from the field of artificial intelligence. Classical AI research—particularly the branch known as connectionism, the basis for artificial neural networks—adheres to the outside-in, tabula rasa model. This prevailing view was perhaps most explicitly promoted in the 20th century by Alan Turing, the great pioneer of mind modeling: "Presumably the child brain is something like a notebook as one buys it from the stationer's," he wrote.

Artificial neural networks built to "write" inputs

© 2023 Scientific American

Imagining the Road Ahead

An experiment demonstrates that distinct sets of neurons fire—each set in a different order—when a rat is planning whether to take the left or right route to receive a reward.

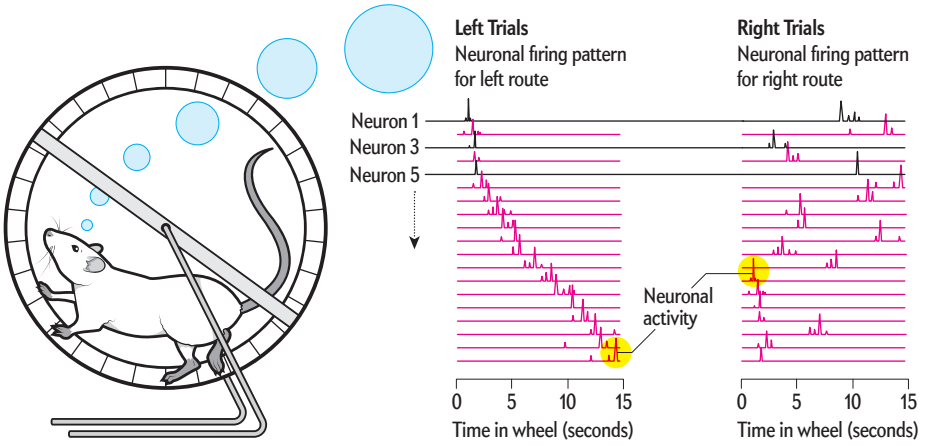


Experimental Setup

A running wheel is located at the entrance to a maze with two route options, both of which lead to a reward. The rat is free to choose a path through the maze after a 15-second run on the wheel. Neuronal firing patterns are recorded during both maze and wheel activity.

Results

Neuronal activity while the rat was running in the wheel predicted the direction it would take in the maze many seconds later, as if the animal were imagining the path to come. The "left trials" panel represents a sequence of neuronal firing that differs from the one for the right trials. When the "left" pattern occurred while the rat was in the wheel, it took the left route in the maze moments later.



onto a neural circuit often fail because each new input inevitably modifies the circuit's connections and dynamics. The circuit is said to exhibit plasticity. But there is a pitfall. While constantly adjusting the connections in its networks when learning, the AI system, at an unpredictable point, can erase all stored memories—a bug known as catastrophic interference, an event a real brain never experiences.

The inside-out model, in contrast, suggests that self-organized brain networks should resist such perturbations. Yet they should also exhibit plasticity selectively when needed. The way the brain strikes this balance relates to vast differences in the connection strength of different groups of neurons. Connections among neurons exist on a continuum. Most neurons are only weakly connected to others, but a smaller subset retains robust links. The strongly connected minority is always on the alert. It fires rapidly, shares information readily within its own group, and stubbornly resists any modifications to the neurons' circuitry. Because of the multitude of connections and their high communication speeds, these elite subnetworks, sometimes described as a "rich club," remain well informed about neuronal events throughout the brain.

The hard-working rich club makes up roughly 20 percent of the overall population of neurons, but it is in charge of nearly half of the brain's activity. In contrast to the rich club, most of the brain's neurons—the neural "poor club"—tend to fire slowly and are weakly connected to other neurons. But they are also highly plastic and able to physically alter the connection points between neurons, known as synapses.

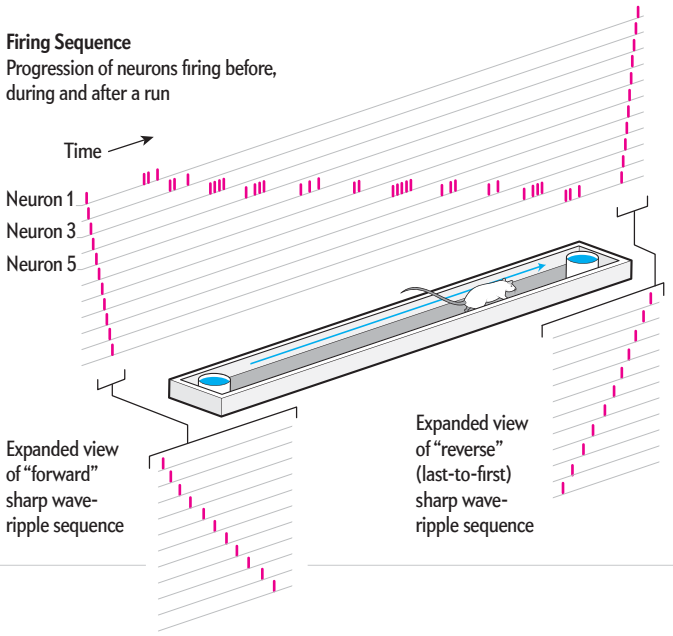
Both rich and poor clubs are key for maintaining brain dynamics. Members of the ever ready rich club fire similarly in response to diverse experiences. They offer quick, good-enough solutions under most conditions. We make good guesses about the unknown not because we remember it but because our brains always make a surmise about a new, unfamiliar event. Nothing is completely novel to the brain because it always relates the new to the old. It generalizes. Even an inexperienced brain has a vast reservoir of neuronal trajectories at the ready, offering opportunities to match events in the world to preexisting brain patterns without requiring substantial reconfiguring of connections. A brain that remakes itself constantly would be unable to adapt

Graphic by Brown Bird Design

© 2023 Scientific American

Rehearsal and Playback

A group of neurons fire before, during and after a rat runs a lap on an elevated track. Neurons firing rapidly at the beginning and end of the run (*insets*) are the same ones active during the run, constituting either a rehearsal or a playback (the prerun sequence in reverse) of the rat's trajectory. These early and late events are known as sharp wave ripples and enable a mental process that selects and remembers an optimal path.


Firing Sequence
Progression of neurons firing before, during and after a run
Time
Neuron 1
Neuron 3
Neuron 5
Expanded view of "forward" sharp wave-ripple sequence
Expanded view of "reverse" (last-to-first) sharp wave-ripple sequence

quickly to fast-changing events in the outside world.

But there also is a critical role for the plastic, slow-firing-rate neurons. These neurons come into play when something of importance to the organism is detected and needs to be recorded for future reference. They then go on to mobilize their vast reserve to capture subtle differences between one thing and another by changing the strength of some connections to other neurons. Children learn the meaning of the word "dog" after seeing various kinds of canines. When a youngster sees a sheep for the first time, they may say "dog." Only when the distinction matters—understanding the difference between a pet and livestock—will they learn to differentiate.

COGNITION AS INTERNALIZED ACTION

AS AN EXPERIMENTER, I did not set out to build a theory in opposition to the outside-in framework. Only decades after I started my work studying the self-organization of brain circuits and the rhythmic firing of neuronal populations in the hippocampus did I realize that the brain is more occupied with itself than with what is happening around it. This realization led to a whole new research agenda for my lab. Our experiments, along with findings from other groups, revealed that neurons devote most of their activity to sustaining the brain's perpetually varying internal states rather than being controlled by stimuli impinging on our senses.

During the course of natural selection, organisms adapt to the ecological niches in which they live and learn to predict the likely outcomes of their actions in those niches. As brain complexity increases, more intricate connections and neuronal computations insert themselves between motor outputs and sensory inputs. This investment enables the prediction of planned actions in more complex and changing environments and at lengthy time scales far in the future. More sophisticated brains also organize themselves to allow computations to continue when sensory inputs vanish temporarily and an animal's actions come to a halt. When you close your eyes, you still know where you are because a great deal of what defines "seeing" is rooted in brain activity. This disengaged mode of neuronal activity provides access to an internalized virtual world of vicarious or imagined experience and serves as a gateway to a variety of cognitive processes.

Let me offer an example of such a disengaged mode of brain operation from our work on the brain's temporal lobe, an area that includes the hippocampus, the nearby entorhinal cortex, and related structures involved with multiple aspects of navigation (the tracking of direction, speed, distance traveled, environmental boundaries, and so on).

Our research builds on leading theories of the functions of the hippocampal system, such as the spectacular Nobel-winning discovery by John O'Keefe of University College London. O'Keefe found that firing of an animal's hippocampal neurons during navigation correlates with that animal's spatial location. For that reason, these neurons are known as place cells.

When a rat walks through a maze, distinct assemblies of place cells become active in a sequential chain corresponding to where it is on its journey. From that observation, one can tentatively conclude that continually changing sensory inputs from the environment exercise control over the firing of neurons, in line with the outside-in model.

Yet other experiments, including in humans, show that these same networks are used for our internal worlds that keep track of personal memories, engage in planning and imagine future actions. If cognition is approached from an inside-out perspective, it becomes clear that navigation through either a physical space or a landscape that exists only in the imagination is processed by identical neural mechanisms.

Fifteen years ago my lab set out to explore the mechanisms of spatial navigation and memory in the hippocampus to contrast the outside-in and inside-out frameworks. In 2008 Eva Pastalkova, then a postdoctoral fellow, and I trained rats to alternate between the left and right arms of a maze to find water. At the beginning of each traversal of the maze, the rat was required to run in a wheel for 15 seconds, which helped to ensure that memory alone of the maze routes, and not environmental and body-derived cues, allowed it to choose

Graphic by Brown Bird Design

© 2023 Scientific American

a particular arm of the maze. We reasoned that if hippocampal neurons "represent" places in the maze corridors and the wheel, as predicted by O'Keefe's spatial navigation theory, a few neurons should fire continuously at each spot whether the rat is in the corridors or the wheel. In contrast, if the neurons' firing is generated by internal brain mechanisms that can support both navigation and memory, the duration of neuronal firing should be similar at all locations, including inside the wheel.

The findings of these experiments defied outside-in explanations. Not a single neuron among the hundreds recorded fired continuously throughout the wheel running. Instead many neurons fired transiently one after the other in a continuous sequence.

Obviously these neurons could not be called place cells, because the animal's body was not displaced while at the single location of the running wheel. Moreover, the firing patterns of individual neurons in this neuronal trajectory could not be distinguished from neurons active when the rat was traversing the arms of the maze.

When we sorted individual trials according to the rat's future choice of left or right arms, the neuronal trajectories were uniquely different. The distinct trajectories eliminated the possibility that these neuronal sequences arose from counting steps, estimating muscular effort or some other undetected feedback stimuli from the body. Also, the unique neuronal trajectories allowed us to predict the animal's maze arm choice from the moment it entered the wheel and throughout wheel running, a period in which the rat had to keep in mind the previously visited arm. The animals needed to correctly choose the alternate maze arm each time to get their rewards [*see box on page 85*].

These experiments lead us to the idea that the neuronal algorithms that we can use to walk to the supermarket govern internalized mental travel. Disengaged navigation takes us through the series of events that make up personal recollections, known as episodic memories.

In truth, episodic memories are more than recollections of past events. They also let us look ahead to plan for the future. They function as a kind of "search engine" that allows us to probe both past and future. This realization also presages a broadening in nomenclature. These experiments show that progressions of place cell activity are internally generated as preconfigured sequences selected for each maze corridor. Same mechanism, multiple designations—so they can be termed place cells, memory cells or planning cells, depending on the circumstance.

Further support for the importance of disengaged circuit operations comes from "offline" brain activity when an animal is milling around doing nothing, consuming a reward or just sleeping. As a rat rests in the home cage after a maze exploration, its hippocampus generates brief, self-organized neuronal trajectories. These sharp wave ripples, as they are known, occur in 100-millisecond time windows and reactivate the same neurons that were firing during several seconds of maze running, recapitulating the neuronal sequences that occurred during maze traversals. Sharp wave-ripple sequences help to form our long-term memories and are essential to normal brain functioning. In fact, alteration of sharp wave-ripple events by experimental manipulations or disease results in serious memory impairment [*see box on opposite page*].

Clever experiments performed in human subjects and in animals over the past decade show that the time-compressed ripple events constitute an internalized trial-and-error process that subconsciously creates real or fictive alternatives for making decisions about an optimal strategy, constructing novel inferences and planning ahead for future actions without having to immediately test them by undertaking a real exploit. In this sense, our thoughts and plans are deferred actions, and disengaged brain activity is an active, essential brain operation. In contrast, the outside-in theory does not make any attempt to assign a role to the disengaged brain when it is at rest or even in the midst of sleep.

THE MEANING OF INSIDE OUT

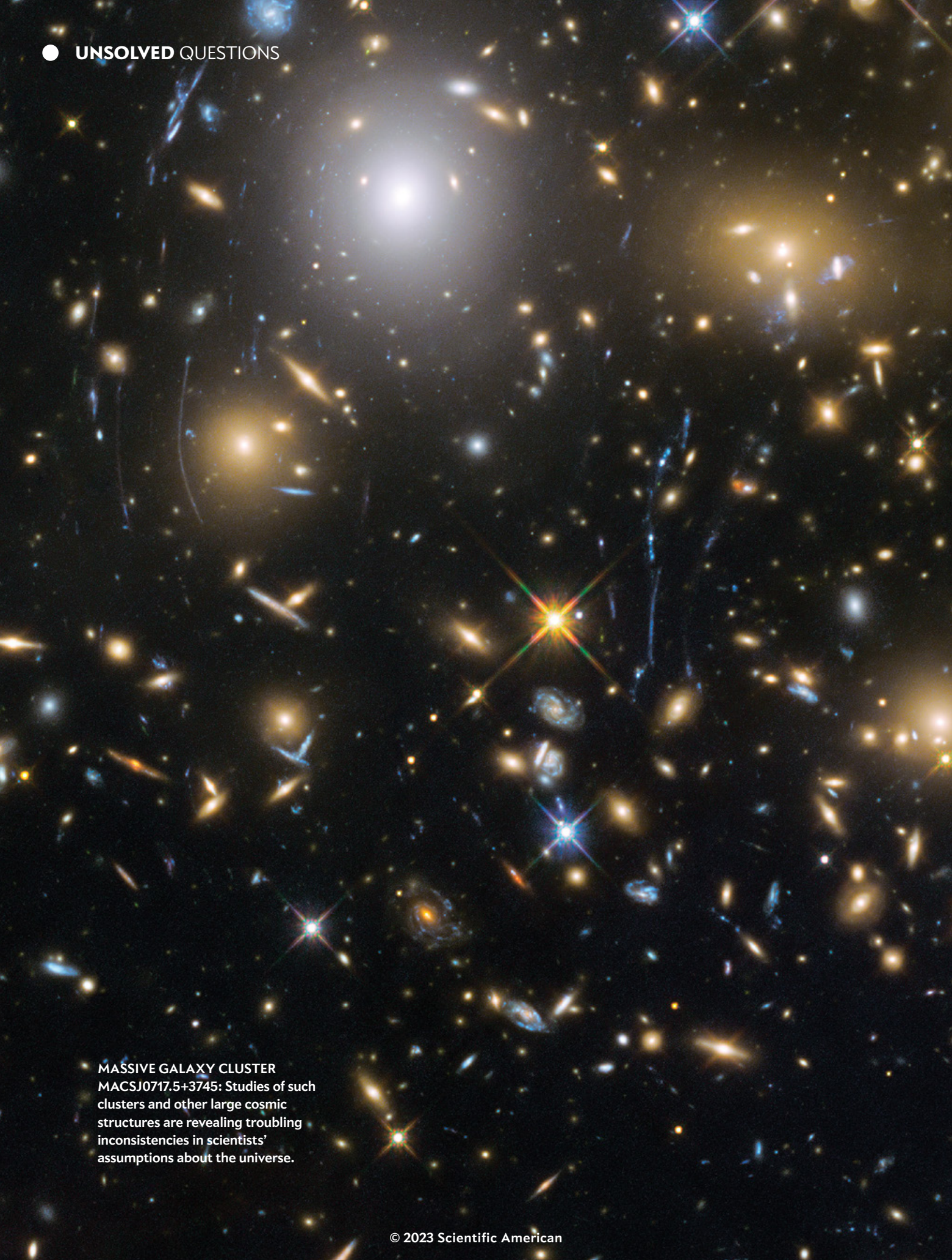
IN ADDITION to its theoretical implications, the inside-out approach has a number of practical applications. It may help in the search to find better diagnostic tools for brain disease. Current terminology often fails to describe accurately underlying biological mechanisms of mental and neurological illnesses. Psychiatrists are aware of the problem but have been hindered by limited understanding of pathological mechanisms and their relation to symptoms and drug responses.

The inside-out theory should also be considered as an alternative to some of the most prevalent connectionist models for conducting AI research. A substitute for them might build models that maintain their own self-organized activity and that learn by "matching" rather than by continual adjustments to their circuitry. Machines constructed this way could disengage their operations from the inputs of electronic sensors and create novel forms of computation that resemble internal cognitive processes.

In real brains, neural processes that operate through disengagement from the senses go hand in hand with mechanisms that promote interactions with the surrounding world. All brains, simple or complex, use the same basic principles. Disengaged neural activity, calibrated simultaneously by outside experience, is the essence of cognition. I wish I had had this knowledge when my smart medical students asked their legitimate questions that I brushed off too quickly. SA

György Buzsáki is a systems neuroscientist whose work has focused on the ways memories form and how brain rhythms segment neural information to support cognition. He was a co-recipient of the 2011 Brain Prize from the Lundbeck Foundation. Buzsáki is author most recently of *The Brain from Inside Out* (Oxford University Press, 2019).

© 2023 Scientific American



MASSIVE GALAXY CLUSTER MACSJ0717.5+3745: Studies of such clusters and other large cosmic structures are revealing troubling inconsistencies in scientists' assumptions about the universe.

© 2023 Scientific American

TWIN TENSIONS

A debate over conflicting measurements of key cosmological properties is poised to shape the next decade of astronomy and astrophysics

By Anil Ananthaswamy

© 2023 Scientific American

HOW FAST IS THE UNIVERSE EXPANDING? HOW MUCH DOES MATTER CLUMP up in our cosmic neighborhood? Scientists use two methods to answer these questions. One involves observing the early cosmos and extrapolating to present times, and the other makes direct observations of the nearby universe. But there is a problem. The two methods consistently yield different answers. The simplest explanation for these discrepancies is merely that our measurements are somehow erroneous, but researchers are increasingly entertaining another, more breathtaking possibility: These twin tensions—between expectation and observation, between the early and late universe—may reflect some deep flaw in the standard model of cosmology, which encapsulates our knowledge and assumptions about the universe. Finding and fixing that flaw could transform our understanding of the cosmos.

One way or another, an answer seems certain to emerge over the coming decade, as new space and terrestrial telescopes give astronomers clearer cosmic views. "Pursuing these tensions is a great way to learn about the universe," says astrophysicist and Nobel laureate Adam Riess of Johns Hopkins University. "They give us the ability to focus our experiments on very specific tests, rather than just making it a general fishing expedition."

These new telescopes, Riess anticipates, are about to usher in the third generation of precision cosmology. The first generation came in the 1990s and early 2000s with the Hubble Space Telescope and with NASA's WMAP satellite, which sharpened our measurements of the universe's oldest light, the cosmic microwave background (CMB). This first generation was also shaped by eight-meter-class telescopes in Chile and the twin 10-meter Keck behemoths in Hawaii.

Collectively, these observatories helped cosmologists formulate their standard model, which holds that the universe is a cocktail of 5 percent ordinary matter, 27 percent dark matter and 68 percent dark energy. This model can account with uncanny accuracy for most of what we observe about galaxies, galaxy clusters, and other large-scale structures and their evolution over cosmic time. Ironically, by its very success, the model highlights what we do not know: the exact nature of 95 percent of the universe.

Driven by even more precise measurements of the CMB from the European Space Agency's Planck satellite and various ground-based telescopes, the second generation of precision cosmology supported the standard model but also brought to light the tensions. The focus shifted to reducing so-called systematics: repeatable errors that creep in because of faults in the design of experiments or equipment.

The third generation is only now starting to take the stage with the successful launch and deep-space deployment of Hubble's successor, the James Webb Space Telescope (JWST). On Earth, radio telescope arrays such as the Simons Observatory in the Atacama Desert in Chile and the CMB-S4, a future assemblage of 21 dishes and half a million cryogenically cooled detectors that will be divided between sites in the Atacama and at the South Pole, should take CMB measurements with Planck-surpassing levels of precision.

The centerpieces of the third generation will be telescopes that stare at wide swaths of the sky. The first of these is likely to be the ESA's 1.2-meter Euclid space telescope, due for launch in 2023. Euclid will study the shapes and distributions of billions of galaxies with a gaze that spans about a third of the sky. Its observations will dovetail with those of NASA's Nancy Grace Roman Space Telescope, a 2.4-meter telescope with a field of view about 100 times bigger than Hubble's, which is slated for launch in the mid-2020s. Finally, when it begins operations around 2024, the ground-based Vera C. Rubin Observatory in Chile will map the entire overhead sky every few nights with its 8.4-meter mirror and a three-billion-pixel camera, the largest ever built for astronomy. "We're not going to be limited by noise and by systematics, because these are independent observatories," says astrophysicist Priyamvada Natarajan of Yale University. "Even if we have a systematic in our framework, we should [be able to] figure it out."

SCALING THE DISTANCE LADDER

RIESS WOULD LIKE to see a resolution of the Hubble tension, which arises from differing estimates of the value of the Hubble constant, H_0—the rate at which the universe is expanding. Riess leads a project called Supernovae, H_0, for the Equation of State of Dark Energy (SH0ES). The goal is to measure H_0, starting with the first rung of the so-called cosmic distance ladder, a hierarchy of methods to gauge ever greater celestial expanses.

The first rung—the one concerning the nearest cosmic objects—relies on determining the distance to special stars called Cepheid variables, which pulsate in proportion to their intrinsic luminosity. The longer the pulsation, the brighter the Cepheid. This relation between variability and luminosity makes Cepheids benchmark "standard candles" for determining distances

NASA/ESA/HST Frontier Fields Team (STScI) *(preceding pages)*

© 2023 Scientific American

A NOCTURNAL VIEW of the South Pole Telescope, one of several radio observatories mapping patterns in the cosmic microwave background.

Danvis Collection/Alamy Stock Photo

around the Milky Way and nearby galaxies. They also form the basis of the cosmic distance ladder's second rung, in which astronomers gauge distances to more remote galaxies by comparing Cepheid-derived estimates with those from another, more powerful set of standard candles called type Ia (pronounced "one A") supernovae, or SNe Ia.

Ascending further, astronomers locate SNe Ia in even more far-flung galaxies, using them to establish a relation between distance and a galaxy's redshift, a measure of how fast it is moving away from us. The result is an estimate of H_0.

In December 2021, Riess says, "after a couple of years of taking a deep dive on the subject," the SH0ES team and the Pantheon+ team, which has compiled a large data set of type Ia supernovae, announced the results of nearly 70 different analyses of their combined data. The data included observations of Cepheid variables in 37 host galaxies that contained 42 type Ia supernovae, more than double the number of supernovae studied by SH0ES in 2016. Riess and his co-authors suspect this latest study represents the Hubble's last stand, the outer limits of that hallowed telescope's ability to help them climb higher up the cosmic scale. The set of supernovae now includes "all suitable SNe Ia—of which we are aware—observed between 1980 and 2021" in the nearby universe. In their analysis, H_0 comes out to be 73.04 ± 1.04 kilometers per second per megaparsec.

That number is way off the value obtained by an entirely different method that looks at the other end of cosmic history—the so-called epoch of recombination, when the universe became transparent to light, about 380,000 years after the big bang. The light from this epoch, now stretched to microwave wavelengths because of the universe's subsequent expansion, is detectable as the all-pervading cosmic microwave background. Tiny fluctuations in temperature and polarization of the CMB capture an important signal: the distance a sound wave travels from almost the beginning of the universe to the epoch of recombination.

This length is a useful metric for precision cosmology, and scientists can use it to estimate the value of H_0 by extrapolating to the present-day universe using the standard LCDM model. (L stands for lambda, or dark energy, and CDM for cold dark matter; cold refers to the assumption that dark matter particles are relatively slow-moving.) An analysis published in 2021 combined data from the Planck satellite and two ground-based instruments, the Atacama Cosmology Telescope and the South Pole Telescope, to arrive at an H_0 of 67.49 ± 0.53.

The discrepancy between the two estimates has a statistical significance of five sigma, meaning there is only about a one-in-a-million chance of its being a statistical fluke. "It's certainly at the level that people should take seriously," Riess says. "And they have."

HOW CLUMPY IS THE COSMOS?

THE OTHER TENSION that researchers are starting to take seriously concerns a cosmic parameter called S_8, which depends on the density of matter in the universe and the extent to which it is

© 2023 Scientific American

clumped up rather than evenly distributed. Estimates of S_8 also involve, on one end, measurements of the CMB and, on the other, measurements of the local universe. The CMB-derived value of S_8 in the early universe, extrapolated using LCDM, generates a present-day value of about 0.834.

The local universe measurements of S_8 involve a host of different methods. Among the most stringent are so-called weak gravitational lensing observations, which measure how the average shape of millions of galaxies across large patches of the sky is distorted by the gravitational influence of intervening concentrations of dark and normal matter. Astronomers used the latest data from the Kilo-Degree Survey, which more than doubled its sky coverage from 350 to 777 square degrees of the sky (the full moon, by comparison, spans a mere half a degree), and estimated S_8 as about 0.759. The tension between the early- and late-universe estimates of S_8 has grown from 2.5 sigma in 2019 to three sigma now (or a one-in-740 chance of its being a fluke). "This tension isn't going away," says astronomer Hendrik Hildebrandt of Ruhr University Bochum in Germany. "It has hardened."

There is yet another way to arrive at the value of S_8: by counting the number of the most massive galaxy clusters in some volume of space. Astronomers can do that directly—for example, by using gravitational lensing. They can also count clusters by studying their imprint on the CMB, thanks to something called the Sunyaev-Zeldovich effect. (This effect causes CMB photons to scatter off the hot electrons in clusters of galaxies, creating shadows in the CMB that are proportional to the mass of the cluster.)

A detailed 2019 study that used data from the South Pole Telescope estimated S_8 as 0.749—again, way off from the CMB+LCDM-based estimates. These numbers could be reconciled if the estimates of the masses of these clusters were wrong by about 40 to 50 percent, Natarajan says, although she thinks such substantial revisions are unlikely. "We are not that badly off in the measurement game," she says. "So that's another kind of internal inconsistency, another anomaly pointing to something else."

> "We don't know why dark energy is the dominant component of the universe today."
> —David Spergel *Princeton University*

BREAKING THE TENSIONS

GIVEN THESE TENSIONS, it is no surprise that cosmologists are anxiously awaiting fresh data from the new generation of observatories. For instance, David Spergel of Princeton University is eager for astronomers to use JWST to study the brightest of the so-called red-giant-branch stars. These stars have a well-known luminosity and can be used as standard candles to measure galactic distances—an independent rung on the cosmic ladder, if you will. In 2019 Wendy Freedman of the University of Chicago and her colleagues used this technique to estimate H_0, finding that their value sits smack in the middle of the early- and late-universe estimates. "The error bars on the current tip of the red-giant-branch data are such that they're consistent with both possibilities," Spergel says. Astronomers are also planning to use JWST to recalibrate the Cepheids surveyed by Hubble, and separately the telescope will help create another rung for the distance ladder by targeting Mira stars (which, like Cepheids, have a luminosity-periodicity relation useful for cosmic cartography).

JWST might resolve or strengthen the H_0 tension, and the wide-field survey data from the Euclid, Roman and Rubin observatories could do the same for the S_8 tension by helping astronomers study the clustering and clumping of matter. The sheer amount of data expected from this trio of telescopes will reduce S_8 error bars enormously. "The statistics are going to go through the roof," Natarajan says.

Meanwhile theoreticians are already having a field day with the twin tensions. "This is a playground for theorists," Riess says. "You throw in some actual observed tensions, and they are having more fun than we are."

The most recent theoretical idea to receive a great deal of interest is something called early dark energy (EDE). In the canonical LCDM model, dark energy started dominating the universe relatively late in cosmic history, only about five billion years ago. But, Spergel says, "we don't know why dark energy is the dominant component of the universe today. Because we don't know why it's important today, it could have also been important early on." That is partly the rationale for invoking dark energy's effects much earlier, before the epoch of recombination. Even if dark energy was just 10 percent of the universe's energy budget during those times, that would be enough to accelerate the early phases of cosmic expansion, causing recombination to occur sooner and shrinking the distance traversed by primordial sound waves. The net effect would be to ease the H_0 tension.

"What I find most interesting about these models is that they can be wrong," Spergel says. Cosmologists' EDE models make predictions about the resulting EDE-modulated patterns in the photons of the CMB. In February 2022 Silvia Galli, a member of the Planck collaboration at the Sorbonne University in Paris, and her colleagues published an analysis of observations from Planck and ground-based CMB telescopes, suggesting that they collectively favor EDE over LCDM by a statistical smidgen. Confirming or refuting this tentative result will require more and better data—which could come soon from upcoming observations by the same ground-based CMB telescopes. But even if EDE models prove to be better fits and fix the H_0 tension, they do little to alleviate the tension from S_8.

Potential fixes for S_8 exhibit a similarly vexing lack of overlap with H_0. In March 2022 Guillermo Franco Abellán of the University of Amsterdam and his colleagues published a study in *Physical Review D* showing that the S_8 tension could be eased by the hypothetical decay of cold dark matter particles into one particle with mass and one "warm" massless particle. This mechanism would lower the value of S_8 arising from CMB-based extrapolations, bringing it more in line with the late-universe measurements. Unfortunately, it doesn't solve the H_0 tension. "It seems like a robust pattern: whatever model you come up with that solves the H_0 tension makes the S_8 tension worse, and the other way around," Hildebrandt says. "There are a few models that at least don't make the other tension worse, but [they] also don't improve it a lot."

© 2023 Scientific American

AN ARTIST'S CONCEPTION of the James Webb Space Telescope, which has just begun to perform breakthrough studies of both the early and the current universe.

NASA/GSFC/CIL/Adriana Manrique Gutierrez

"WE ARE MISSING SOMETHING"

SPERGEL FORESEES a few scenarios that could be possible once fresh data arrive. First, the new CMB data might turn out to be consistent with early dark energy, resolving the H_0 tension, and the upcoming survey telescope observations might separately ease the S_8 tension. That would be a win for early dark energy models—and would constitute a major shift in our understanding of the opening chapters of cosmic history. It's also possible that both H_0 and S_8 tensions will resolve in favor of LCDM. This would be a win for the cosmic standard model and a possibly bittersweet victory for cosmologists hoping for paradigm-shifting breakthroughs. Of course, it might turn out that neither tension is resolved. "Outcome three would be both tensions become increasingly significant as the data improve—and early dark energy isn't the answer," Spergel says. Then, LCDM would presumably have to be reworked differently, although how is unclear.

Natarajan thinks that the tensions and discrepancies are probably telling us that LCDM is merely an "effective theory," a technical term meaning that it accurately explains a certain subset of the current compendium of cosmic observations. "Perhaps what's really happening is that there is an underlying, more complex theory," she says. "And LCDM is this [effective] theory, which seems to have most of the key ingredients. For the level of observational probes we had previously, that effective theory was sufficient." But times change, and the data deluge from precision cosmology's third generation of powerful observatories may demand more creative and elaborate theories.

Theorists, of course, are happy to oblige. For instance, Spergel speculates that if early dark energy could interact with dark matter (in LCDM, dark energy and dark matter do not interact), the arrangement could suppress the fluctuations of matter in the early universe in ways that would resolve the S_8 tension while simultaneously taking care of the H_0 tension. "It makes the models more baroque," Spergel says, "but maybe that's what nature will demand."

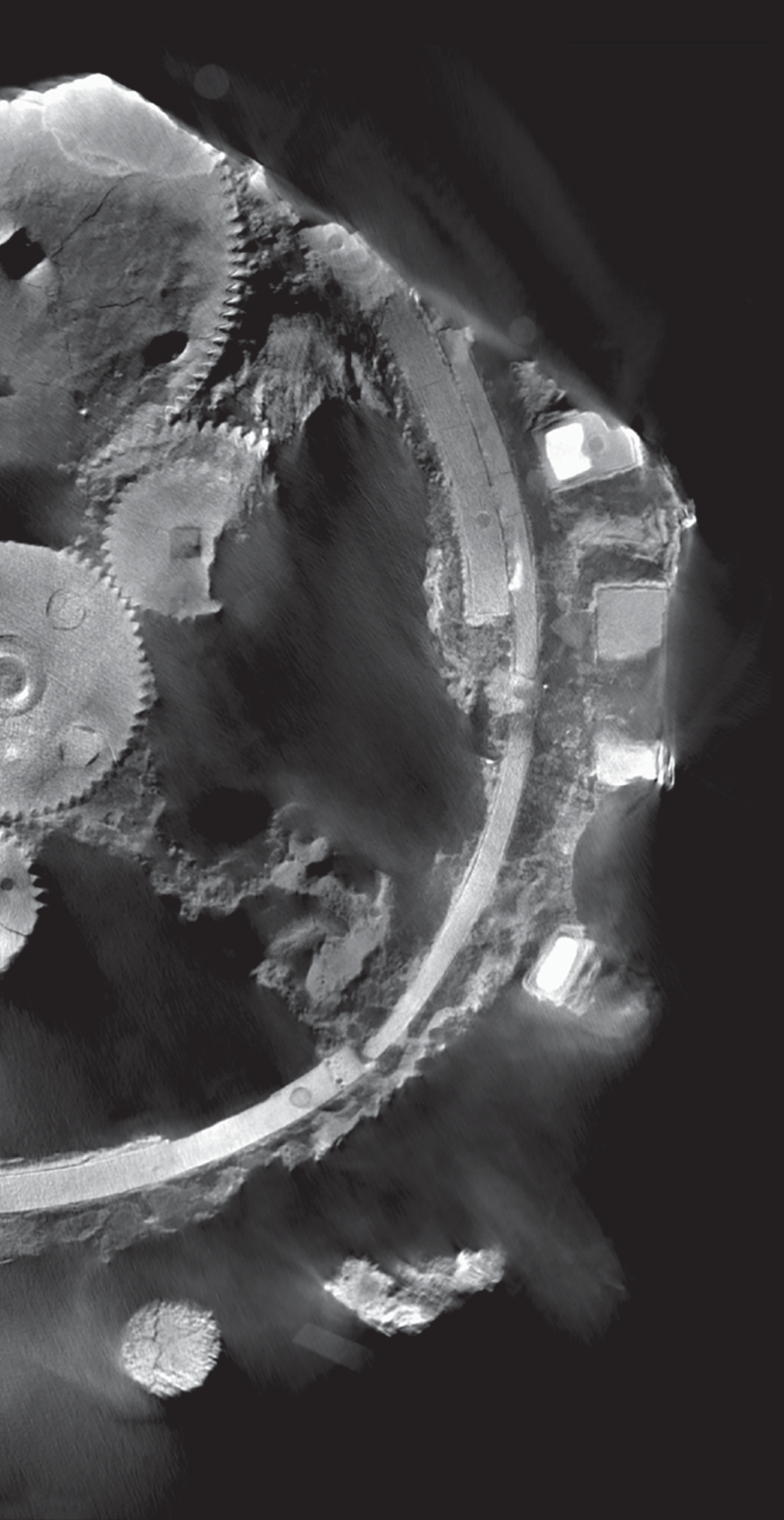
As an observational astronomer, Hildebrandt is circumspect. "If there was a convincing model that beautifully solves these two tensions, we'd already have the next standard model," he says. "That we're instead still talking about these tensions and scratching our heads is just reflecting the fact that we don't have such a model yet." Riess agrees. "After all, this is a problem of using a model based on an understanding of physics and the universe that is about 95 percent incomplete, in terms of the nature of dark matter and dark energy," he says. "It wouldn't be crazy to think that we are missing something." SA

Anil Ananthaswamy is author of *The Edge of Physics* (Houghton Mifflin Harcourt, 2010), *The Man Who Wasn't There* (Dutton, 2015), and, most recently, *Through Two Doors at Once: The Elegant Experiment That Captures the Enigma of Our Quantum Reality* (Dutton, 2018).

© 2023 Scientific American

WONDER OF THE ANCIENT WORLD

© 2023 Scientific American



Scientists have revealed intriguing details about the complex gearing of the Antikythera mechanism of ancient Greece

By Tony Freeth

OUTSIDE/INSIDE: The main drive wheel of the Antikythera mechanism, as seen in a photograph of its present state (*left*) and in an x-ray CT scan.

© 2023 Scientific American

IN 1900 DIVER ELIAS STADIATIS, CLAD IN A COPPER AND BRASS HELMET AND A HEAVY canvas suit, emerged from the sea shaking in fear and mumbling about a "heap of dead naked people." He was among a group of Greek divers from the eastern Mediterranean island of Symi who were searching for natural sponges. They had sheltered from a violent storm near the tiny island of Antikythera, between Crete and mainland Greece. When the storm subsided, they dived for sponges and chanced on a shipwreck full of Greek treasures—the most significant wreck from the ancient world to have been found up to that point. The "dead naked people" were marble sculptures scattered on the seafloor, along with many other artifacts. Soon after, their discovery prompted the first major underwater archaeological dig in history.

One object recovered from the site, a lump the size of a large dictionary, initially escaped notice amid more exciting finds. Months later, however, at the National Archaeological Museum in Athens, the lump broke apart, revealing bronze precision gearwheels the size of coins. According to historical knowledge at the time, gears like these should not have appeared in ancient Greece, or anywhere else in the world, until many centuries after the shipwreck. The find generated huge controversy.

The lump is known as the Antikythera mechanism, an extraordinary object that has befuddled historians and scientists for more than 120 years. Over the decades the original mass split into 82 fragments, leaving a fiendishly difficult jigsaw puzzle for researchers to put back together. The device appears to be a geared astronomical calculation machine of immense complexity. Today we have a reasonable grasp of some of its workings, but there are still unsolved mysteries. We know it is at least as old as the shipwreck it was found in, which has been dated to between 60 and 70 B.C.E., but other evidence suggests it may have been made around 200 B.C.E.

In March 2021 my group at University College London, known as the UCL Antikythera Research Team, published a new analysis of the machine. The team on that research included me (a mathematician and filmmaker); Adam Wojcik (a materials scientist); Lindsay MacDonald (an imaging scientist); the late Myrto Georgakopoulou (an archaeometallurgist); and two graduate students, David Higgon (a horologist) and Aris Dacanalis (a physicist). Our paper posited a new explanation for the gearing on the front of the mechanism, where the evidence had previously been unresolved. Our findings gave us an even better appreciation for the sophistication of the device—an understanding that challenges many of our preconceptions about the technological capabilities of the ancient Greeks.

ANCIENT ASTRONOMY

WE KNOW THE GREEKS of that era were accomplished naked-eye astronomers. They viewed the night sky from a geocentric perspective—every night, as Earth turned on its axis, they saw the dome of stars rotating. The stars' relative positions remained unchanged, so the Greeks called them the "fixed stars." These early astronomers also saw bodies moving against the background of stars: the moon goes through a rotation against the stars every 27.3 days; the sun takes a year.

The other moving bodies are the planets, named "wanderers" by the Greeks because of their erratic motions. They were the deepest problem for astronomy at the time. Scientists wondered what they were and noticed that sometimes the wanderers move in the same direction as the sun—in "prograde" motion—then come to a stop and reverse direction to move in "retrograde." After a while they reach another stationary point and resume prograde motion again. These rotations are called the synodic cycles of the planets—their cycles relative to the sun. The seemingly strange reversals happen because, as we know now, the planets orbit the sun—not, as the ancient Greeks believed, Earth.

In modern terms, all the moving astronomical bodies have orbits close to the plane of Earth's motion around the sun—the so-called ecliptic—meaning that they all follow much the same path through the stars. Predicting the positions of the planets along the ecliptic was very difficult for early astronomers. This task, it turns out, was one of the primary functions of the Antikythera mechanism. Another function was to track the positions of the sun and moon, which also have variable motions against the stars.

Much of the mechanism's design relies on wisdom from earlier Middle Eastern scientists. Astronomy in particular went through a transformation during the first millennium B.C.E. in Babylon and Uruk (both in modern-day Iraq). The Babylonians recorded the daily positions of the astronomical bodies on clay tablets, which revealed that the sun, moon and planets moved in repeating cycles—a fact that was critical for making predictions. The moon, for instance, goes through 254 cycles against the backdrop of the stars every 19 years—an example of a so-called period relation. The Antikythera mechanism's design uses several of the Babylonian period relations.

© 2005 National Archaeological Museum in Athens (*preceding pages, left side*);
© 2005 Nikon X-Tek Systems (*preceding pages, right side*)

© 2023 Scientific American

A Celestial Calculation Machine

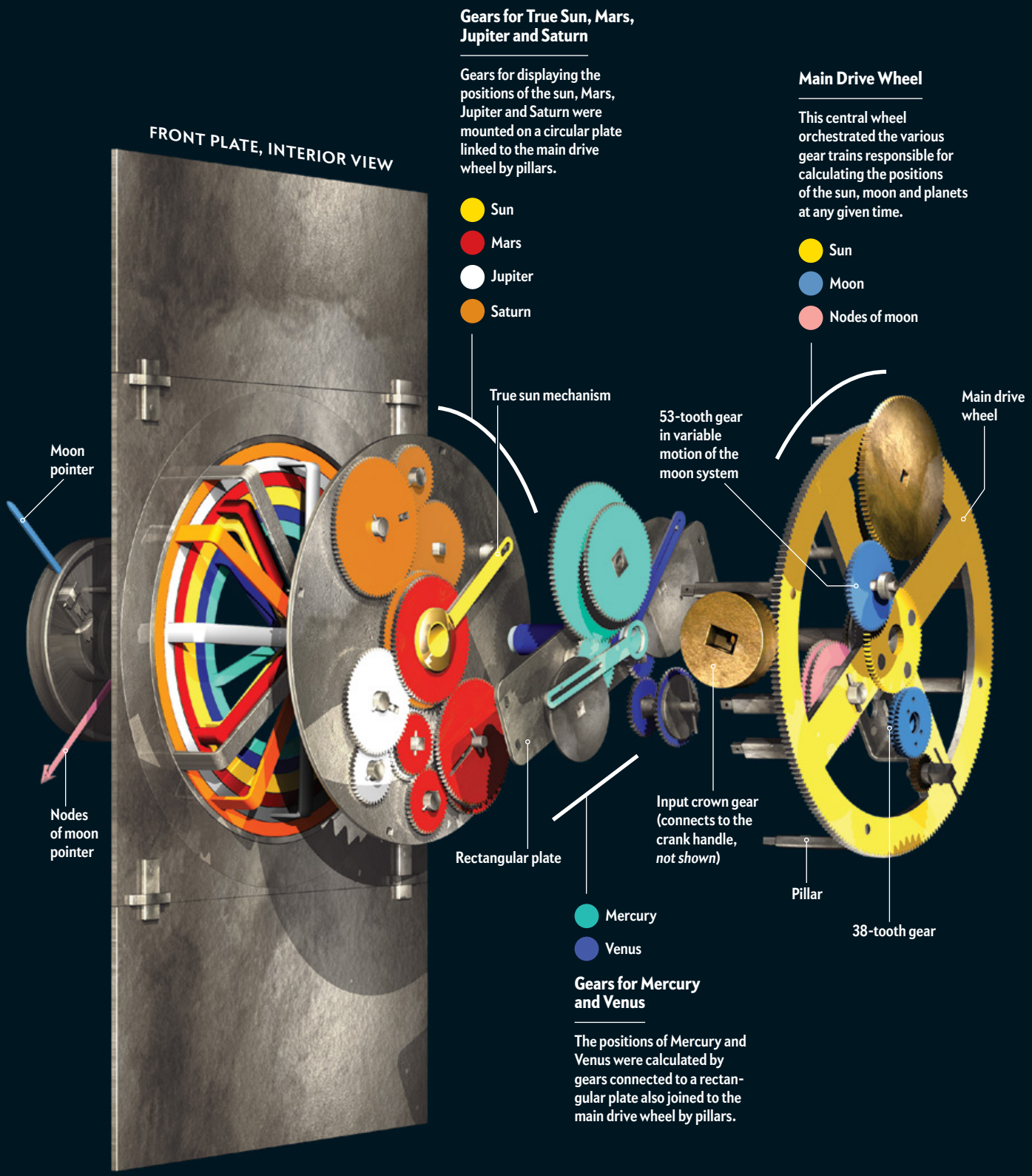
Discovered more than a century ago in a shipwreck, the Antikythera mechanism is the most technologically complex object ever found from the ancient world. Likely dating back to between 205 and 60 B.C.E., the bronze device contains dozens of small gears with teeth about a millimeter long that were used to predict the positions of the sun, moon and planets at any chosen time. A study by the University College London Antikythera Research Team proposed a new model of how the gearing on the front of the machine worked.

Front: Inside the device a "main drive wheel" turned all the gearing, which moved pointers and concentric rings displaying the positions of different celestial bodies. Small spheres showed the positions of the sun and moon and the phase of the moon. Colored beads marked the locations of the planets along the ecliptic, the plane of the solar system.

Back: The rear face contained two large dials, as well as some smaller ones. The large top dial was a calendar representing the Metonic cycle, a 19-year period over which 235 moon phases recur. The large lower dial was the 223-month saros dial, which predicted the dates of solar and lunar eclipses.

Graphic by Tony Freeth

© 2023 Scientific American

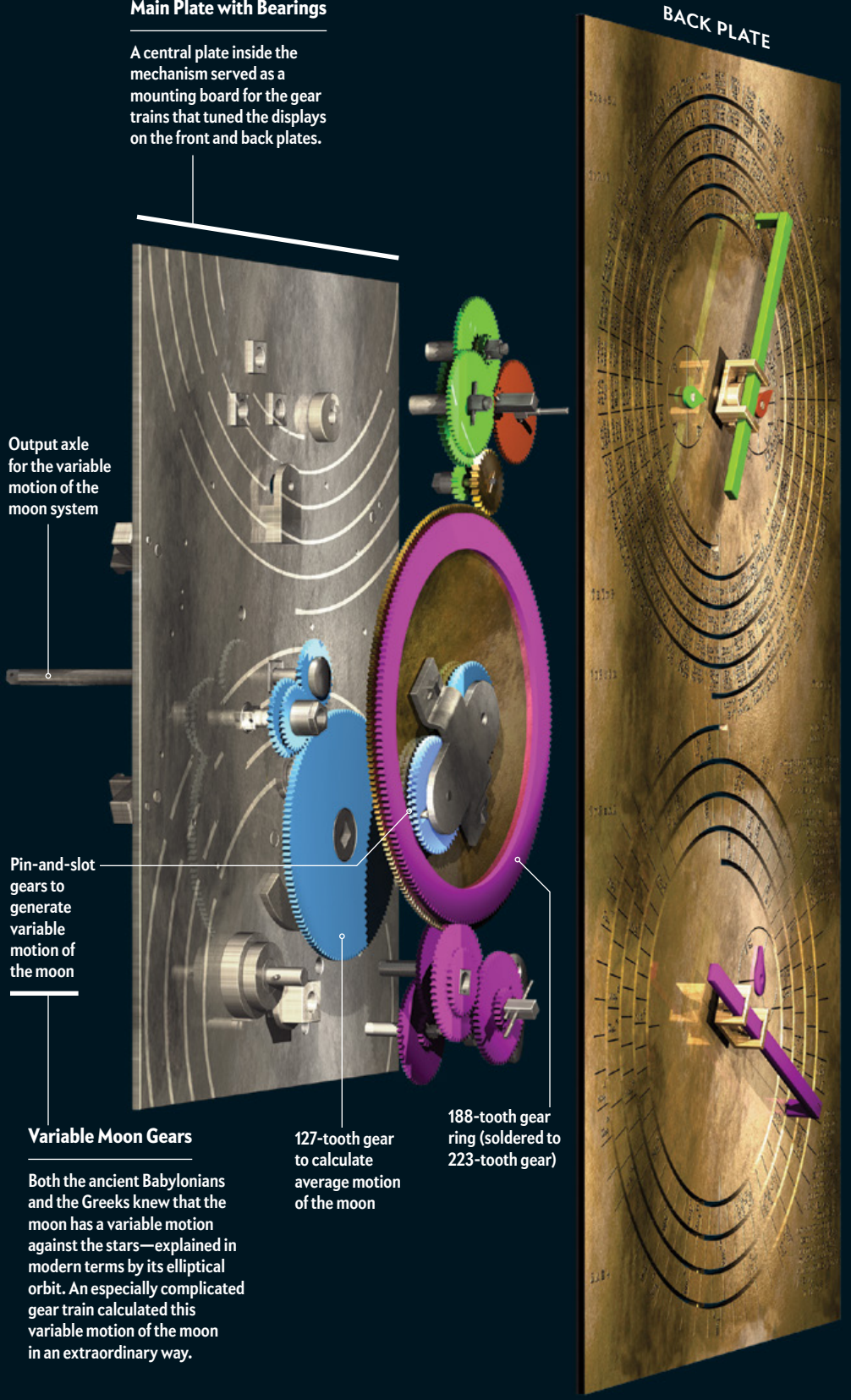

FRONT PLATE, INTERIOR VIEW
Gears for True Sun, Mars, Jupiter and Saturn
Gears for displaying the positions of the sun, Mars, Jupiter and Saturn were mounted on a circular plate linked to the main drive wheel by pillars.
Sun
Mars
Jupiter
Saturn
Main Drive Wheel
This central wheel orchestrated the various gear trains responsible for calculating the positions of the sun, moon and planets at any given time.
Sun
Moon
Nodes of moon
True sun mechanism
53-tooth gear in variable motion of the moon system
Main drive wheel
Moon pointer
Nodes of moon pointer
Rectangular plate
Input crown gear (connects to the crank handle, not shown)
Pillar
38-tooth gear
Mercury
Venus
Gears for Mercury and Venus
The positions of Mercury and Venus were calculated by gears connected to a rectangular plate also joined to the main drive wheel by pillars.

Graphic by Tony Freeth and Jen Christiansen, Model by UCL Antikythera Research Team

© 2023 Scientific American

INSIDE THE MACHINE

The new model of the Antikythera mechanism proposes a total of 69 gears, forming a shockingly complex astronomical calculation tool. Most of this complexity was hidden in the innards of the device, where trains of gears worked together on different calculations, and some gears served double duty for multiple purposes. From the outside, a user could turn a calendar dial to a desired point in the past, present or future, and the gears inside would move pointers and rings on the surface display to reveal celestial predictions.


Main Plate with Bearings
A central plate inside the mechanism served as a mounting board for the gear trains that tuned the displays on the front and back plates.
BACK PLATE
Output axle for the variable motion of the moon system
Pin-and-slot gears to generate variable motion of the moon
Variable Moon Gears
Both the ancient Babylonians and the Greeks knew that the moon has a variable motion against the stars—explained in modern terms by its elliptical orbit. An especially complicated gear train calculated this variable motion of the moon in an extraordinary way.
127-tooth gear to calculate average motion of the moon
188-tooth gear ring (soldered to 223-tooth gear)
Metonic and Kallippic calendar
Olympiad calendar
Saros and exeligmos calendar
Back Dials
The upper back dial system was a Metonic/Kallippic calendar that reconciled the lunar month with the solar year. It also included a smaller dial showing the four-year Olympiad cycle of the Panhellenic Games, commonly used to mark time. The lower back dial system was a saros/exeligmos calendar that predicted solar and lunar eclipses according to the 223-month saros cycle. It was indexed to inscriptions on the back plate that describe the characteristics of the predicted eclipses.

© 2023 Scientific American

One of the central researchers in the early years of Antikythera research was German philologist Albert Rehm, the first person to understand the mechanism as a calculating machine. Between 1905 and 1906 he made crucial discoveries that he recorded in his unpublished research notes. He found, for instance, the number 19 inscribed on one of the surviving Antikythera fragments. This figure was a reference to the 19-year period relation of the moon known as the Metonic cycle, named after Greek astronomer Meton but discovered much earlier by the Babylonians. On the same fragment, Rehm found the numbers 76, a Greek refinement of the 19-year cycle, and 223, for the number of lunar months in a Babylonian eclipse-prediction cycle called the saros cycle. These repeating astronomical cycles were the driving force behind Babylonian predictive astronomy.

The second key figure in the history of Antikythera research was British physicist turned historian of science Derek J. de Solla Price. In 1974, after 20 years of research, he published an important paper, "Gears from the Greeks." It referred to remarkable quotations by Roman lawyer, orator and politician Cicero (106–43 B.C.E.). One of these described a machine made by mathematician and inventor Archimedes (circa 287–212 B.C.E.) "on which were delineated the motions of the sun and moon and of those five stars which are called wanderers … (the five planets) … Archimedes … had thought out a way to represent accurately by a single device for turning the globe those various and divergent movements with their different rates of speed." This machine sounds just like the Antikythera mechanism. The passage suggests that Archimedes, although he lived before we believe the device was built, might have founded the tradition that led to the Antikythera mechanism. It may well be that the Antikythera mechanism was based on a design by Archimedes.

The Antikythera mechanism, with its precision gears bearing teeth about a millimeter long, is completely unlike anything else from the ancient world.

FIENDISHLY COMPLEX

FOR DECADES RESEARCHERS were stuck trying to decipher the workings of the device by looking at the surface of its disintegrating fragments. In the early 1970s they finally got to peek inside. Price worked with Greek radiologist Charalambos Karakalos to obtain x-ray scans of the fragments. To their astonishment, the researchers found 30 distinct gears: 27 in the largest fragment and one each in three others. Karakalos, with his wife, Emily, was able to estimate the tooth counts of the gearwheels for the first time, a critical step in understanding what the mechanism calculated. The machine was looking more complicated than anyone had conceived.

The x-ray scans were two-dimensional, meaning that the structure of the gearing appeared flattened, and they revealed only partial pictures of most of the gears. Scientists could only infer the number of teeth on many of the gears. Despite these shortcomings, Price identified a gear train—a set of linked gears—that calculated the average position of the moon on any specific date by using its period relation of 254 sidereal rotations in 19 years. Driven by a prominent feature on the front of the mechanism called the main drive wheel, this gear train starts with a 38-tooth gear (two times 19, as a gear with just 19 teeth would be a bit too small). This 38-tooth gear drives (via some other gears) a 127-tooth gear (half of 254; the full number would require too large a gear).

It seems that the device could be used to predict the positions of the sun, moon and planets on any specific day in the past or future. The maker of the machine would have had to calibrate it with the known positions of these bodies. A user could then simply turn a crank to the desired time frame to see astronomical predictions. The mechanism displayed positions, for instance, on a "zodiac dial" on the front of the mechanism, where the ecliptic was divided into a dozen 30-degree sections representing the constellations of the zodiac. Based on the x-ray data, Price developed a complete model of all the gearing on the device.

Price's model was my introduction to the Antikythera mechanism. My first paper, in fact, "Challenging the Classic Research," was a comprehensive demolition of most of Price's proposed gearing structure for the machine. Nevertheless, Price correctly determined the relative positions of the major fragments and defined the overall architecture of the machine, with date and zodiac dials at the front and two large dial systems at the back. Price's achievements were a significant step in decoding the Antikythera mystery.

A third key figure in the history of Antikythera research is Michael Wright, a former curator of mechanical engineering at London's Science Museum. In collaboration with Australian professor of computer science Alan G. Bromley, Wright carried out a second x-ray study of the mechanism in 1990 using an early 3-D x-ray technique called linear tomography. Bromley died before this work bore fruit, but Wright was persistent, making important advances, for example, in identifying the crucial tooth counts of the gears and in understanding the upper dial on the back of the device.

In 2000 I proposed the third x-ray study, which was carried out in 2005 by a team of academics from England and Greece in collaboration with the National Archaeological Museum in Athens. X-Tek Systems (now owned by Nikon) developed a prototype x-ray machine to take high-resolution 3-D x-ray images using microfocus x-ray computed tomography (x-ray CT). Hewlett-Packard used a brilliant digital imaging technique called polynomial texture mapping for enhancing surface details.

The new data surprised us. The first major breakthrough was my discovery that the mechanism predicted eclipses in addition to the motions of the astronomical bodies. This finding was connected to the inscription Rehm had found that mentioned the 223-month saros eclipse cycle. The new x-rays revealed a large, 223-tooth gear at the rear of the mechanism that turns a pointer around a dial that spirals out, making four turns in total that are divided into 223 sections, for 223 months. Named after the customary name of the Babylonian eclipse cycle, the saros dial predicts which months will feature eclipses, along with characteristics of each eclipse as described by inscriptions in the mechanism. The finding revealed an impressive new feature of the device, but

© 2023 Scientific American

it left a massive problem: a group of four gears lying within the circumference of the large gear that appeared to have no function.

It took months to understand these gears. When I did, the results were astonishing. These gears turned out to calculate the variable motion of the moon in a very beautiful way. In modern terms, the moon has variable motion because it has an elliptical orbit: when it is farther from Earth, it moves more slowly against the stars; when it is closer, it moves more quickly. The moon's orbit, however, is not fixed in space: the whole orbit rotates in a period of just under nine years. The ancient Greeks did not know about elliptical orbits, but they explained the moon's subtle motion by combining two circular motions in what is called an epicyclic theory.

I figured out how the mechanism calculated the epicyclic theory by building on a remarkable observation by Wright. He had studied two of the four mysterious gears at the back of the mechanism. He saw that one of them has a pin on its face that engages with a slot on the other gear. It might seem to be a useless arrangement because the gears will surely just turn together at the same rate. But Wright noticed that the gears turn on different axes separated by just over a millimeter, meaning that the system generates variable motion. All these details appear in the x-ray CT scan. The axes of the gears are not fixed—they are mounted epicyclically on the large, 223-tooth gear.

Wright discarded the idea that these gears calculated the moon's variable motion because in his model, the 223-tooth gear turned much too fast for it to make sense. But in my model, the 223-tooth gear rotates very slowly to turn the pointer for the saros dial. Calculating the epicyclic theory of the moon with epicyclic pin-and-slot gears in this subtle and indirect way was an extraordinary conception by the ancient Greeks. This ingenuity reinforces the idea that the machine was designed by Archimedes. This research on the back dials and gearing completed our understanding of the back of the mechanism, reconciling all the evidence up to that point. My colleagues and I published our findings in 2006 in *Nature*. The other side of the device, however, was still very much a mystery.

HIDDEN MESSAGE: X-ray CT scans made in 2005 revealed previously unseen inscriptions on the Antikythera mechanism, including a list of planetary cycles on the front cover (*shown here*) and a "user's manual" on the back cover.

THE FRONT OF THE MECHANISM

THE MOST PROMINENT FEATURE of the front of the largest fragment is the main drive wheel, which was designed to rotate once a year. It is not a flat disk like most of the other gears; this one has four spokes and is covered in puzzling features. The spokes show evidence that they held bearings: there are circular holes in them for turning axles. The outer edge of the gear contains a ring of pillars—little fingers that stick up perpendicularly, with shoulders and pierced ends that were clearly intended to carry plates. Four short pillars held a rectangular plate, and four long pillars held a circular plate.

Following Price, Wright proposed that an extensive epicyclic system—the two-circles idea the Greeks used to explain the odd reversing motions of the planets—had been mounted on the main drive wheel. Wright even constructed an actual model gearing system in brass to show how it worked. In 2002 he published a groundbreaking planetarium model for the Antikythera mechanism that displayed all five planets known in the ancient world. (The discovery of Uranus and Neptune in the 18th and 19th centuries, respectively, required the advent of telescopes.) Wright showed that the epicyclic theories could be translated into epicyclic gear trains with pin-and-slot mechanisms to display the planets' variable motions.

When I first saw Wright's model, I was shocked by its mechanical complexity. It even featured eight coaxial outputs—tubes all centered on a single axis—that brought information to the front display of the device. Was it really plausible that the ancient Greeks could build such an advanced system? I now believe that Wright's conception of coaxial outputs must be correct, but his gearing system does not match the economy and ingenuity of the known gear trains. The challenge our UCL team faced was to reconcile Wright's coaxial outputs with what we knew about the rest of the device.

One crucial clue came from the 2005 x-ray CT study. In addition to showing the gears in three dimensions, these scans made an unexpected revelation—thousands of new text characters hidden inside the fragments and unread for more than 2,000 years. In his research notes from 1905 to 1906, Rehm proposed that the positions of the sun and planets were displayed in a concentric system of rings. The mechanism originally had two covers—front and back—that protected the displays and included extensive inscriptions. The back-cover inscription, revealed in the 2005 scans, was a user manual for the device. In 2016 Alexander Jones, a professor of the history of astronomy at New York University,

© 2005 Nikon X-Tek Systems

© 2023 Scientific American

discovered definitive evidence for Rehm's idea within this inscription: a detailed description of how the sun and planets were displayed in rings, with marker beads to show their positions.

Any model for the workings of the mechanism should match this description—an explanation literally inscribed onto the back cover of the device describing how the sun and planets were displayed. Yet previous models had failed to incorporate this ring system because of a technical problem that we could not solve. Wright had discovered that the device used a semisilvered ball to show the phase of the moon, which it calculated mechanically by subtracting an input for the sun from an input for the moon. But such a process appeared to be incompatible with a ring system for displaying the planets because the outputs for Mercury and Venus prevented the moon-phase device from accessing the input from the sun gear system. In 2018 Higgon, one of the graduate students on our UCL team, came up with a surprisingly simple idea that neatly fixed this technical problem and explained a mysterious pierced block on one of the spokes of the main drive wheel. This block could transmit the "mean sun" rotation (as opposed to the variable "true sun" rotation) directly to the moon-phase device. This setup enabled a ring system for the front of the Antikythera mechanism that fully reflected the description in the back-cover inscription.

In trying to decipher the front of the device, it was imperative to identify the planetary cycles built into the mechanism because they define how the gear trains calculated planetary positions. Earlier research assumed that they would be based on the planetary period relations derived by the Babylonians. But in 2016 Jones made a discovery that forced us to discard that assumption.

The x-ray CT of the front-cover inscription shows it is divided into sections for each of the five planets. In the Venus section, Jones found the number 462; in the Saturn section, he found the number 442. These numbers were astonishing. No previous research had suggested that ancient astronomers knew them. In fact, they represent more accurate period relations than the ones found by the Babylonians. It seems that the makers of the Antikythera device discovered their own improved period relations for two of the planets: 289 synodic cycles in 462 years for Venus and 427 synodic cycles in 442 years for Saturn.

Jones never figured out how the ancient Greeks derived both these periods. We set out to try ourselves. Dacanalis, our other UCL graduate student, assembled a comprehensive list of the planetary period relations and their estimated errors from Babylonian astronomy. Could combinations of these earlier relations be the key to the more accurate Antikythera period relations? Eventually we found a process, developed by philosopher Parmenides of Elea (sixth to fifth century B.C.E.) and reported by Plato (fifth to fourth century B.C.E.), for combining known period relations to get better ones.

We proposed that any method the Antikythera creators used would have required three criteria: accuracy, factorizability and economy. The method must be accurate to match the known period relations for Venus and Saturn, and it must be factorizable so the planets could be calculated with gears small enough to fit into the mechanism. To make the system economical, different planets could share gears if their period relations shared prime factors, reducing the number of gears needed. Such economy is a key feature of the surviving gear trains. Based on these criteria, our team derived the periods 462 and 442 using the idea from Parmenides and employed the same methods to discover the missing periods for the other planets where the inscriptions were lost or damaged.

Armed with the period relations for the planets, we could now understand how to fit the gear trains for the planets into the tight spaces available. For Mercury and Venus, we theorized economical five-gear mechanisms with pin-and-slot devices, similar to Wright's mechanisms for these planets. We found strong supporting evidence for our reconstruction in one four-centimeter-diameter fragment. Inside this piece, the x-ray CT shows a disk attached to a 63-tooth gear, which turns in a D-shaped plate. The number 63 shares the prime factors 3 and 7 with 462 (the Venus period). A gear train using the 63-tooth gear could be designed to match a bearing on one of the spokes of the main drive wheel. A similar design for Mercury matches the features on the opposite spoke. These observations gave us great confidence that we were on the right track for Mercury and Venus.

For the other known planets—Mars, Jupiter and Saturn—our team conceived of very compact systems to fit the available space. These designs were a radical departure from Wright's systems for these planets. Working independently, Christián C. Carman of the National University of Quilmes in Argentina and I had shown that the subtle indirect gearing system for the variable motion of the moon could be adapted for these planets. Our UCL team proved that these gearing systems could be extended to incorporate the new period relations for the planets. This system allowed the Antikythera makers to mount several gears on the same plate and design them to precisely match the period relations.

These economical seven-gear trains could intricately interleave between the plates on the pillars of the main drive wheel so that their outputs conformed to the customary cosmological order of the celestial bodies—moon, Mercury, Venus, sun, Mars, Jupiter and Saturn—that determines the layout of the ring system. The dimensions of the available spaces between the plates were exactly right to accommodate these systems, with some spare capacity and some evidence still unexplained.

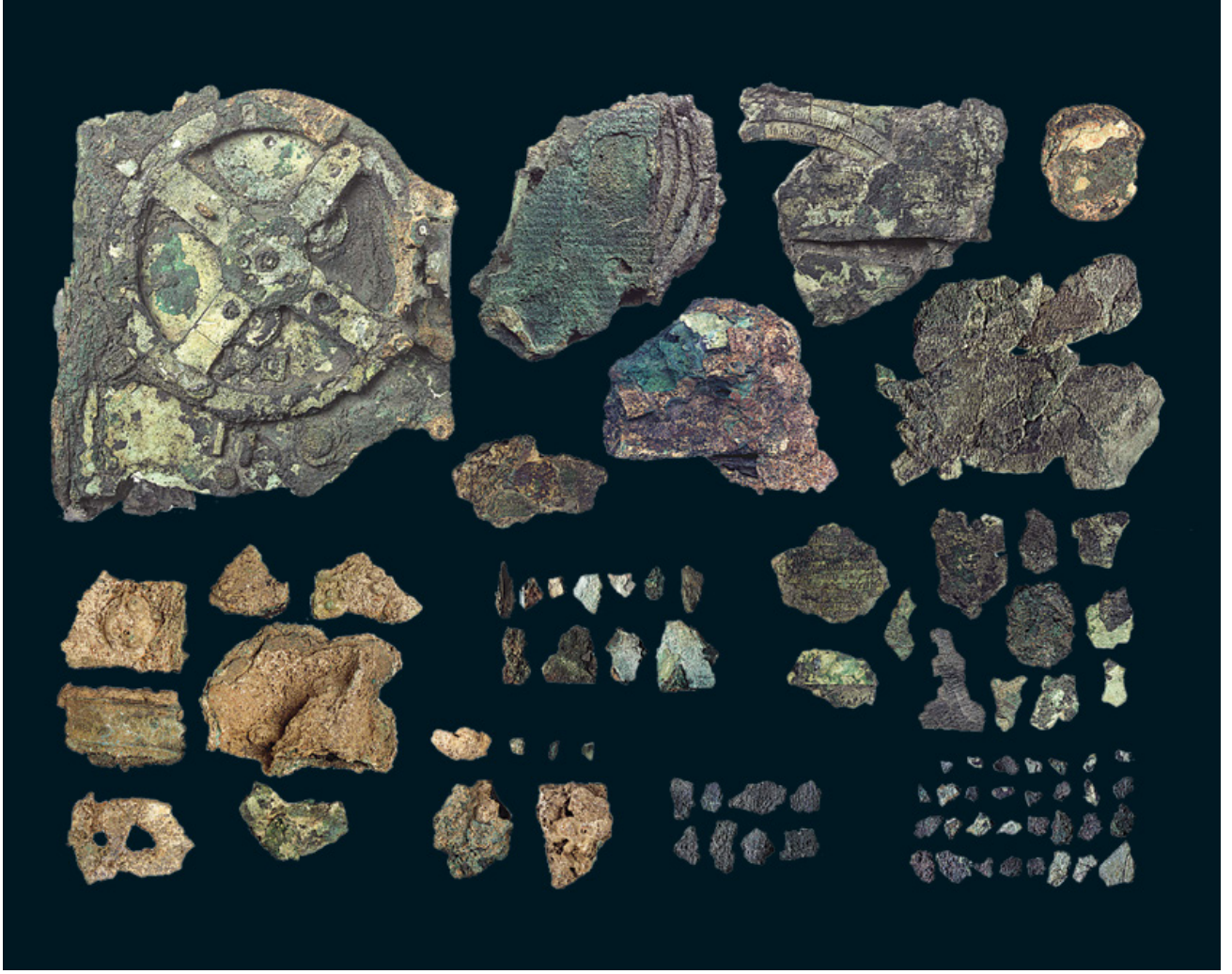
We added a mechanism for the variable motion of the sun and an epicyclic mechanism for calculating the "nodes" of the moon—the points at which the moon's orbit cuts through the plane of the ecliptic, making an eclipse possible. Eclipses happen only when the sun is close to one of these nodes during a full or new moon. Medieval and renaissance astronomers called a double-ended pointer for the nodes of the moon a "dragon hand." The epicyclic gearing for this dragon hand also exactly explained a prominent bearing on one of the spokes that had previously appeared to have no function. We had finally explained all the features on the main drive wheel; we published our findings in March 2021 in *Scientific Reports*.

A BEAUTIFUL CONCEPTION

WE NOW UNDERSTOOD HOW the front display matched the description in the back-cover user's manual, with the sun and planets shown by marker beads on concentric rings. The front cover also displayed the moon's phase, position and age (the number of days from a new moon), and the dragon hand that showed eclipse years and seasons.

With the concentric rings for the planets, we realized that we could now make sense of the front-cover inscription as well. This writing is a formulaic list of the synodic events of each planet

© 2023 Scientific American



FRAGMENTS: Over the years the original mass of the Antikythera mechanism has split into 82 pieces. Figuring out how they all fit together has been a challenging puzzle for researchers. The largest fragment (*top left*) holds the main drive wheel.

(such as its conjunctions with the sun and its stationary points) and the intervals in days between them. On the back plate, the eclipse inscriptions are indexed to markings on the saros dial. On the front plate, inscriptions about the risings and settings of stars are indexed to the zodiac dial. Our insight was that the inscriptions on the front could refer to index letters on the planetary rings: if the sun pointer is at one of these letters, then the corresponding inscription entry describes the number of days to the next synodic event. Because the left-hand side of the inscription, where we would expect these index letters to be, is missing, we cannot prove the hypothesis—but it is a compelling explanation.

The device is unique among discoveries from its time. It single-handedly rewrites our knowledge of the technology of the ancient Greeks. We knew they were highly capable—they built the Parthenon and the Lighthouse of Alexandria even earlier than the Antikythera mechanism. They had plumbing and used steam to operate equipment. But before the discovery of the Antikythera mechanism, ancient Greek gears were thought to be restricted to crude wheels in windmills and water mills. Aside from this discovery, the first precision-geared mechanism known is a relatively simple—yet impressive for the time—geared sundial and calendar of Byzantine origin dating to about C.E. 600. It was not until the 14th century that scientists created the first sophisticated astronomical clocks. The Antikythera mechanism, with its precision gears bearing teeth about a millimeter long, is completely unlike anything else from the ancient world.

Why did it take centuries for scientists to reinvent anything as sophisticated as the Antikythera device, and why haven't archaeologists uncovered more such mechanisms? We have strong reasons to believe this object can't have been the only model of its kind—there must have been precursors to its development. But bronze was a very valuable metal, and when an object like this stopped working, it probably would have been melted down for its materials. Shipwrecks may be the best prospects for finding more of them. As for why the technology was seemingly lost for so long before being redeveloped, who knows? There are many gaps in the historical record, and future discoveries may well surprise us.

With the Antikythera mechanism, we are clearly not at the end of our story. We believe our work is a significant advance, but there are still mysteries to be solved. The UCL Antikythera Research Team is not certain that our reconstruction is entirely correct because of the huge loss of evidence. It is very hard to match all of the surviving information. Regardless, we can now see more clearly than ever what a towering achievement this object represents. SA

Tony Freeth is a member of the University College London Antikythera Research Team. A mathematician and an award-winning filmmaker, Freeth has conducted research on the Antikythera mechanism and promoted it through films and presentations since 2000.

© 2005 National Archaeological Museum in Athens

© 2023 Scientific American

OUR INNER UNIVERSES

Reality is constructed by the brain, and no two brains are exactly alike

By Anil K. Seth

Illustration by Brook VanDevelder

"We do not see things as they are, we see them as we are."
—from *Seduction of the Minotaur,* by Anaïs Nin (1961)

© 2023 Scientific American

© 2023 Scientific American

ON THE 10TH OF APRIL 2019 POPE FRANCIS, PRESIDENT SALVA KIIR OF South Sudan and former rebel leader Riek Machar sat down together for dinner at the Vatican. They ate in silence, the start of a two-day retreat aimed at reconciliation from a civil war that had killed some 400,000 people since 2013. At about the same time in my laboratory at the University of Sussex in England, Ph.D. student Alberto Mariola was starting to work on an experiment in which volunteers experience being in a room they believe is there but is not. In psychiatry clinics across the globe, people arrive complaining that things no longer seem "real" to them, whether it is the world around them or their own selves. In the fractured societies in which we live, what is real—and what is not—seems to be increasingly up for grabs. Warring sides may experience and believe in different realities. Perhaps eating together in silence can help because it offers a small slice of reality that can be agreed on, a stable platform on which to build further understanding.

We need not look to war and psychosis to find radically different inner universes. In 2015 a badly exposed photograph of a dress tore across the Internet, dividing the world into those who saw it as blue and black (me included) and those who saw it as white and gold (half my lab). Those who saw it one way were so convinced they were right—that the dress truly was blue and black or white and gold—that they found it almost impossible to believe that others might perceive it differently.

We all know that our perceptual systems are easy to fool. The popularity of visual illusions is testament to this phenomenon. Things seem to be one way, and they are revealed to be another: two lines appear to be different lengths, but when measured they are exactly the same; we see movement in an image we know to be still. The story usually told about illusions is that they exploit quirks in the circuitry of perception, so that what we perceive deviates from what is there. Implicit in this story, however, is the assumption that a properly functioning perceptual system will render to our consciousness things precisely as they are.

The deeper truth is that perception is never a direct window onto an objective reality. All our perceptions are active constructions, brain-based best guesses at the nature of a world that is forever obscured behind a sensory veil. Visual illusions are fractures in the Matrix, fleeting glimpses into this deeper truth.

Take, for example, the experience of color—say, the bright red of the coffee mug on my desk. The mug really does seem to be red: its redness seems as real as its roundness and its solidity. These features of my experience seem to be truly existent properties of the world, detected by our senses and revealed to our mind through the complex mechanisms of perception.

Yet we have known since Isaac Newton that colors do not exist out there in the world. Instead they are cooked up by the brain from mixtures of different wavelengths of colorless electromagnetic radiation. Colors are a clever trick that evolution has hit on to help the brain keep track of surfaces under changing lighting conditions. And we humans can sense only a tiny slice of the full electromagnetic spectrum, nestled between the lows of infrared and the highs of ultraviolet. Every color we perceive, every part of the totality of each of our visual worlds, comes from this thin slice of reality.

Just knowing this is enough to tell us that perceptual experience cannot be a comprehensive representation of an external objective world. It is both less than that and more than that. The reality we experience—the way things *seem*—is not a direct reflection of what is actually out there. It is a clever construction by the brain, for the brain. And if my brain is different from your brain, my reality may be different from yours, too.

THE PREDICTIVE BRAIN

IN PLATO'S *Allegory of the Cave,* prisoners are chained to a blank wall all their lives, so that they see only the play of shadows cast by objects passing by a fire behind them, and they give the shadows names because for them the shadows are what is real. A thousand years later, but still a thousand years ago, Arabian scholar Ibn al-Haytham wrote that perception, in the here and now, depends on processes of "judgment and inference" rather than involving direct access to an objective reality. Hundreds of years later again Immanuel Kant realized that the chaos of unrestricted sensory data would always remain meaningless without being given structure by preexisting conceptions or "beliefs," which for him included a priori frameworks such as space and time. Kant's term "noumenon" refers to a "thing in itself"—*Ding an sich*—an objective reality that will always be inaccessible to human perception.

Today these ideas have gained a new momentum through an influential collection of theories that turn on the idea that the brain is a kind of prediction machine and that perception of the world—and of the self within it—is a process of brain-based predictions about the causes of sensory signals.

These new theories are usually traced to German physicist and physiologist Hermann von Helmholtz, who in the late 19th century proposed that perception is a process of unconscious inference. Toward the end of the 20th century Helmholtz's notion was taken up by cognitive scientists and artificial-intelligence researchers, who reformulated it in terms of what is now generally known as predictive coding or predictive processing.

The central idea of predictive perception is that the brain is attempting to figure out what is out there in the world (or in here, in the body) by continually making and updating best guesses about the causes of its sensory inputs. It forms these best guesses by combining prior expectations or "beliefs" about the world, together with incoming sensory data, in a way that takes into account how reliable the sensory signals are. Scientists usually conceive of this process as a form of Bayesian inference, a framework that specifies how to update beliefs or best guesses with new data when

© 2023 Scientific American

both are laden with uncertainty.

In theories of predictive perception, the brain approximates this kind of Bayesian inference by continually generating predictions about sensory signals and comparing these predictions with the sensory signals that arrive at the eyes and the ears (and the nose and the fingertips and all the other sensory surfaces on the outside and inside of the body). The differences between predicted and actual sensory signals give rise to so-called prediction errors, which are used by the brain to update its predictions, readying it for the next round of sensory inputs. By striving to minimize sensory-prediction errors everywhere and all the time, the brain implements approximate Bayesian inference, and the resulting Bayesian best guess is what we perceive.

To understand how dramatically this perspective shifts our intuitions about the neurological basis of perception, it is helpful to think in terms of bottom-up and top-down directions of signal flow in the brain. If we assume that perception is a direct window onto an external reality, then it is natural to think that the content of perception is carried by bottom-up signals—those that flow from the sensory surfaces inward. Top-down signals might contextualize or finesse what is perceived, but nothing more. Call this the "how things seem" view because it seems as if the world is revealing itself to us directly through our senses.

POORLY EXPOSED photograph of a dress appears blue and black to some people, white and gold to others.

The prediction machine scenario is very different. Here the heavy lifting of perception is performed by the top-down signals that convey perceptual predictions, with the bottom-up sensory flow serving only to calibrate these predictions, keeping them yoked, in some appropriate way, to their causes in the world. In this view, our perceptions come from the inside out just as much as, if not more than, from the outside in. Rather than being a passive registration of an external objective reality, perception emerges as a process of active construction—a controlled hallucination, as it has come to be known.

Why controlled hallucination? People tend to think of hallucination as a kind of false perception, in clear contrast to veridical, true-to-reality, normal perception. The prediction machine view suggests instead a continuity between hallucination and normal perception. Both depend on an interaction between top-down, brain-based predictions and bottom-up sensory data, but during hallucinations, sensory signals no longer keep these top-down predictions appropriately tied to their causes in the world. What we call hallucination, then, is just a form of uncontrolled perception, just as normal perception is a controlled form of hallucination.

This view of perception does not mean that nothing is real. Writing in the 17th century, English philosopher John Locke made an influential distinction between "primary" and "secondary" qualities. Primary qualities of an object, such as solidity and occupancy of space, exist independently of a perceiver. Secondary qualities, in contrast, exist only in relation to a perceiver—color is a good example. This distinction explains why conceiving of perception as controlled hallucination does not mean it is okay to jump in front of a bus. The bus has primary qualities of solidity and space occupancy that exist independently of our perceptual machinery and that can do us injury. It is the way in which the bus appears to us that is a controlled hallucination, not the bus itself.

TRIPPING IN THE LAB

A GROWING BODY of evidence supports the idea that perception is controlled hallucination, at least in its broad outlines. A 2015 study by Christoph Teufel of Cardiff University in Wales and his colleagues offers a striking example. In this study, the ability to recognize so-called two-tone images was evaluated in patients with early-stage psychosis who were prone to hallucinations.

Take a look at the photograph on the next page—a sample of a two-tone image. Probably all you will see is a bunch of black-and-white splotches. Now look at the image on page 109. Then have another look at the first photo; it ought to look rather different. Where previously there was a splotchy mess, there are now distinct objects, and something is happening.

What I find remarkable about this exercise is that in your second examination of the first image, the sensory signals arriving at your eyes have not changed at all from the first time you saw it. All that has changed are your brain's predictions about the causes of these sensory signals. You have acquired a new high-level perceptual expectation, and this changes what you consciously see.

If you show people many two-tone images, each followed by the full picture, they might subsequently be able to identify a good proportion of two-tone images, though not all of them. In Teufel's study, people with early-stage psychosis were better at recognizing two-tone images after having seen the full image than were healthy control subjects. In other words, being hallucination-prone went along with perceptual priors having a stronger effect on perception. This is exactly what would be expected if hallucinations in psychosis depended on an overweighting of perceptual priors so that they overwhelmed sensory-prediction errors, unmooring perceptual best guesses from their causes in the world.

Recent research has revealed more of this story. In a 2021 study, Biyu He of New York University and her colleagues had neurosurgical patients look at ambiguous images, such as a Necker cube, that constantly flip between two different appearances even though the sensory input remains the same. By analyzing

Swiked.tumblr.com

© 2023 Scientific American

TWO-TONE IMAGE looks like a mess of black-and-white splotches, until you see the full image (*opposite page*).

the signals recorded from within the patients' brains, they discovered that information flowed more strongly in a top-down, "inside-out" direction when the perceived appearance was consistent with the patients' biases, as would be expected if perceptual predictions were strong in this case. And when the perceived appearance was inconsistent with preexisting biases, information flow was stronger in the bottom-up direction, suggesting a "prediction error" signal. This is an exciting new development in mapping the brain basis of controlled hallucinations.

In my lab we have taken a different approach to exploring the nature of perception and hallucination. Rather than looking into the brain directly, we decided to simulate the influence of overactive perceptual priors using a unique virtual-reality setup masterminded by our resident VR guru, Keisuke Suzuki. We call it, with tongue firmly in cheek, the "hallucination machine."

Using a 360-degree camera, we first recorded panoramic video footage of a busy square in the University of Sussex campus on a Tuesday at lunchtime. We then processed the footage through an algorithm based on Google's AI program DeepDream to generate a simulated hallucination. What happens is that the algorithm takes a so-called neural network—one of the workhorses of AI—and runs it backward. The network we used had been trained to recognize objects in images, so if you run it backward, updating the network's input instead of its output, the network effectively projects what it "thinks" is there onto and into the image. Its predictions overwhelm the sensory inputs, tipping the balance of perceptual best guessing toward these predictions. Our particular network was good at classifying different breeds of dogs, so the video became unusually suffused with dog presences.

Many people who have viewed the processed footage through the VR headset have commented that the experience is rather reminiscent not of the hallucinations of psychosis but of the exuberant phenomenology of psychedelic trips.

More recently, we have implemented the hallucination machine in different ways to simulate different kinds of altered visual experience. By extending our algorithm to include two coupled neural networks—a "discriminator network" much like the one in our original study and a "generator network" that has been trained to reproduce ("generate") its input image—we have been able to model different types of hallucination. For example, we have modeled the complex hallucinatory experiences reported by people with Parkinson's disease and some forms of dementia; the patterned, geometric hallucinations that occur after the loss of foveal vision, as happens in Charles Bonnet syndrome; and a range of psychedeliclike hallucinations. We hope that by understanding hallucinations better, we will be able to understand normal experience better, too, because predictive perception is at the root of all our perceptual experience.

THE PERCEPTION OF REALITY

ALTHOUGH THE HALLUCINATION machine is undoubtedly trippy, people who experience it are fully aware that what they are experiencing is not real. Indeed, despite rapid advances in VR technology and computer graphics, no current VR setup delivers an experience sufficiently convincing to be indistinguishable from reality.

This is the challenge we took up when designing a new "substitutional reality" setup at Sussex—the one we were working on when Pope Francis convened the retreat with Salva Kiir and Riek Machar. Our aim was to create a system in which volunteers would experience an environment as being real—and believe it to be real—when in fact it was not real.

The basic idea is simple. We again prerecorded some panoramic video footage, this time of the interior of our VR lab rather than of an outside campus scene. People coming to the lab are invited to sit on a stool in the middle of the room and to put on a VR headset that has a camera attached to the front. They are encouraged to look around the room and to see the room as it actually is, via the camera. At some point, without telling them, we switch the feed so that the headset displays not the live real-world scene but rather the prerecorded panoramic video. Most people in this situation continue to experience what they are seeing as real even though it is now a fake prerecording. (This is actually very tricky to pull off in practice—it requires careful color balancing and alignment to avoid people noticing any difference that would tip them off to the shift.)

I find this result fascinating because it shows that it is possible to have people experience an unreal environment as being fully real. This demonstration alone opens new frontiers for VR research: we can test the limits of what people will experience, and believe, to be real. It also allows us to investigate how experiencing things as being real can affect other aspects of perception. Right now we are running an experiment to find out whether people are worse at detecting unexpected changes in the room when they believe that what they are experiencing is real. If things do turn out this way (the study was heavily delayed by the global pandem-

© 2023 Scientific American

ic, and its results are not yet final), that finding would support the idea that the perception of things as being real itself acts as a high-level prior that can substantively shape our perceptual best guesses, affecting the contents of what we perceive.

PERCEPTUAL SHIFT: Viewing this photograph changes what one consciously sees in the two-tone image (*opposite page*).

THE REALITY OF REALITY

THE IDEA THAT THE WORLD of our experience might not be real is an enduring trope of philosophy and science fiction, as well as of late-night pub discussions. Neo in *The Matrix* takes the red pill, and Morpheus shows him how what he thought was real is an elaborate simulation, while the real Neo lies prone in a human body farm, a brain-in-a-vat power source for a dystopian AI. Philosopher Nick Bostrom of the University of Oxford has famously argued, based largely on statistics, that we are likely to be living inside a computer simulation created in a posthuman age. I disagree with this argument in part because it assumes that consciousness can be simulated—I do not think that this is a safe assumption—but it is thought-provoking nonetheless.

Although these chunky metaphysical topics are fun to chew on, they are probably impossible to resolve. Instead what we have been exploring throughout this article is the relation between appearance and reality in our conscious perceptions, where part of this appearance is the appearance of being real itself.

The central idea here is that perception is a process of active interpretation geared toward adaptive interaction with the world via the body rather than a re-creation of the world within the mind. The contents of our perceptual worlds are controlled hallucinations, brain-based best guesses about the ultimately unknowable causes of sensory signals. For most of us, most of the time, these hallucinations are experienced as real. As Canadian rapper and science communicator Baba Brinkman suggested to me, when we agree about our hallucinations, maybe that's what we call reality.

But we do not always agree, and we do not always experience things as real. People with dissociative psychiatric conditions such as derealization or depersonalization syndrome report that their perceptual worlds, even their own selves, lack a sense of reality. Some kinds of hallucination, various psychedelic hallucinations among them, combine a sense of unreality with perceptual vividness, as does lucid dreaming. People with synesthesia consistently have additional sensory experiences, such as perceiving colors when viewing black letters, which they recognize as not real. Even with normal perception, if you look directly at the sun you will experience the subsequent retinal afterimage as not being real. There are many such ways in which we experience our perceptions as not fully real.

What this means to me is that the property of realness that attends most of our perceptions should not be taken for granted. It is another aspect of the way our brain settles on its Bayesian best guesses about its sensory causes. One might thus ask what purpose it serves. Perhaps the answer is that a perceptual best guess that includes the property of being real is usually more fit for purpose—that is, better able to guide behavior—than one that does not. We will behave more appropriately with respect to a coffee cup, an approaching bus or our partner's mental state when we experience it as really existing.

But there is a trade-off. As illustrated by the dress illusion, when we experience things as being real, we are less able to appreciate that our perceptual worlds may differ from those of others. (A popular explanation for the differing perceptions of the garment holds that people who spend most of their waking hours in daylight see it as white and gold; night owls, who are mainly exposed to artificial light, see it as blue and black.) And even if these differences start out small, they can become entrenched and reinforced as we proceed to harvest information differently, selecting sensory data that are best aligned with our individual emerging models of the world and then updating our perceptual models based on these biased data. We are all familiar with this process from the echo chambers of social media and the newspapers we choose to read. I am suggesting that the same principles apply also at a deeper level, underneath our sociopolitical beliefs, right down to the fabric of our perceptual realities. They may even apply to our perception of being a self—the experience of being me or of being you—because the experience of being a self is itself a perception.

This is why understanding the constructive, creative mechanisms of perception has unexpected social relevance. Perhaps once we can appreciate the diversity of experienced realities scattered among the billions of perceiving brains on this planet, we will find new platforms on which to build a shared understanding and a better future—whether between sides in a civil war, followers of different political parties, or two people sharing a house and faced with washing the dishes. SA

Anil K. Seth is a professor of cognitive and computational neuroscience at the University of Sussex in England. His research focuses on the biological basis of consciousness, and he is author of *Being You—A New Science of Consciousness* (Dutton, 2021).

Richard Armstrong/EyeEm/Getty Images (*opposite page* and *this page*)

© 2023 Scientific American

Long Haulers and Chronic Illnesses

Millions of Americans have long COVID—and we are not prepared to manage the crisis

By Meghan O'Rourke

WHEN THE FIRST WAVE OF CORONAVIRUS INFECTIONS hit the U.S. in March 2020, what kept me up at night was not only the tragedy of the acute crisis but also the idea that we might soon be facing a second crisis—a pandemic of chronic illness triggered by the virus. I had just finished reporting and writing a book about infection-associated syndromes and contested chronic illnesses, long an underresearched and dismissed area of medicine. Medical science has increasingly understood that infections can trigger ongoing physical symptoms in a subset of people, yet the medical establishment has typically ignored the experiences of those people. Such conditions include myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS), so-called chronic Lyme disease, and more.

Sure enough, later that spring a cohort of patients who had caught the coronavirus in March began reporting that they were still not better. In online message boards, patients began sharing stories of what they called long COVID. Groups of "long haulers" banded together to call for more attention and research into their plight.

The clamor, combined with the scope of the problem, had a clear impact on medical attitudes, making long COVID visible in ways that ME/CFS had struggled for decades to become. In a matter of months centers dedicated to treating long COVID sprang up at respected research hospitals, such as the Center for Post-COVID Care at Mount Sinai in New York City. In itself, this is a hopeful development: when I got sick with a similar condition a decade ago, I longed for such a place.

The effects on research, too, have been dramatic, with scientists at numerous academic medical centers working to understand what long COVID is, how to measure it, and how best to treat or manage it. Akiko Iwasaki, an immunologist and head of a laboratory at the Yale School of Medicine, is one of them. "I used to focus mainly on acute infectious diseases, but with long COVID on the rise, a big chunk of my lab now focuses on long COVID and other postacute infection syndromes," she says. David Putrino, director of rehabilitation innovation for the Mount Sinai Health System, says he has seen "a sharp increase in interested researchers," in part because funding agencies such as the National Institutes of Health "have begun allocating increased resources to long COVID."

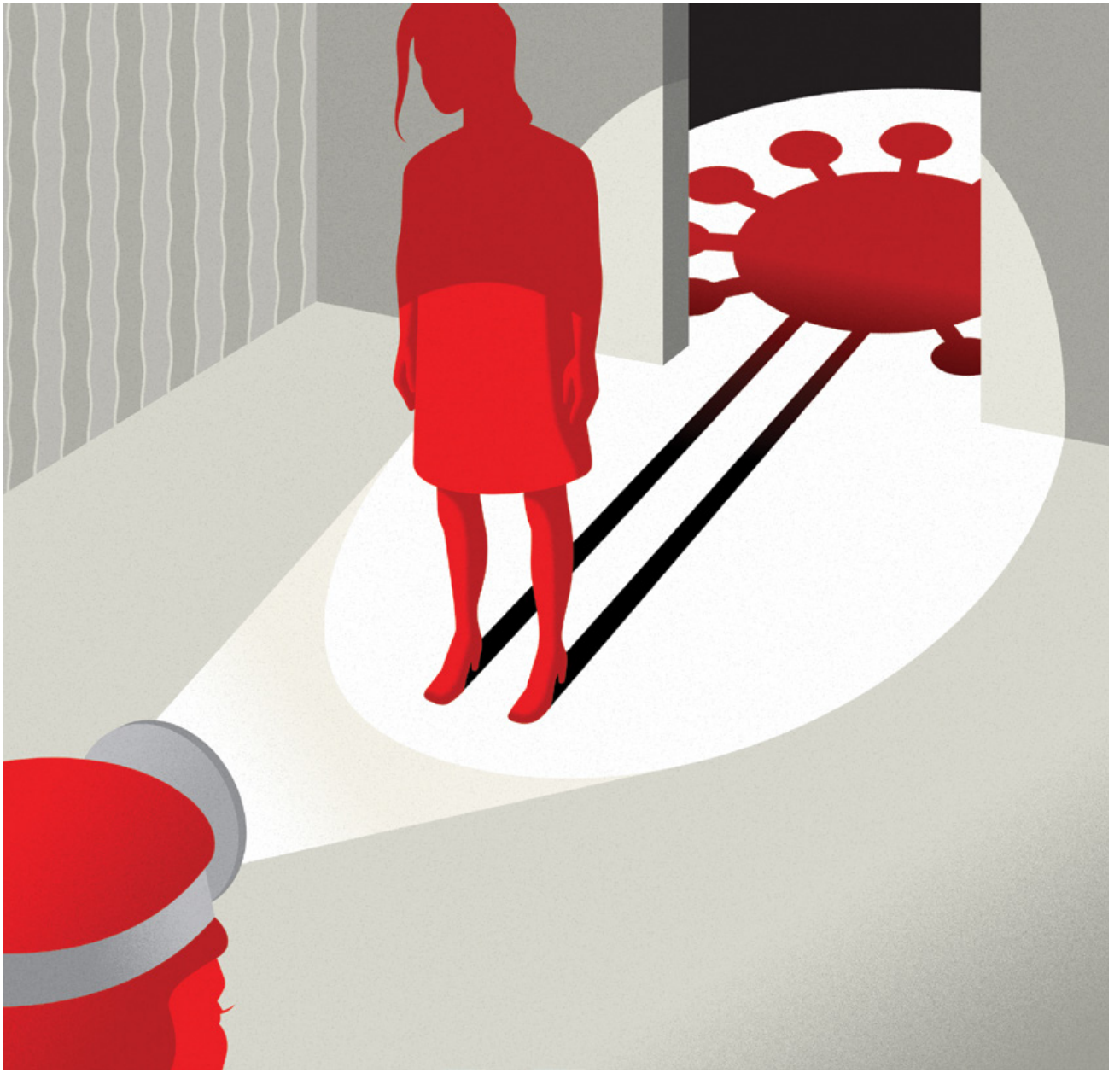
Three years into the pandemic, long COVID remains one of the biggest threats it poses. Early estimates suggested anywhere from 10 to 50 percent of unvaccinated people infected with the virus develop long-term symptoms. Vaccines may reduce the risk by as much as 50 percent, but according to Putrino, they do not eliminate it.

Yet long COVID was rarely discussed in public health messaging during the Delta and Omicron waves; officials focused on acute severe disease and death and largely ignored the debilitating—and life-altering—long-term effects that the virus has on so many people. We've had even fewer conversations about the societal responsibilities we have toward a growing generation of sick people, many of whom are between the ages of 30 and 50.

This lack of concern is even more surprising considering that we still understand little about the condition, including what causes it. Some theories suggest that the virus triggers rampant inflammation or autoimmune disease, others that the virus itself may persist in tissues in the body. What we do know is that millions of people are seeking care for a staggering array of symptoms that include fatigue, brain fog, racing hearts, breathlessness, pain, and more. The task of treating all these patients is exposing some of medicine's enduring weaknesses.

Modern medicine is based on replicability. Since the advent of germ theory in the 19th century, the field has taken an "if you can't measure it, it doesn't exist" view, as Harvard University researcher Susan D. Block put it to me. Medicine has a long history of stigmatizing diseases it does not understand and cannot yet readily measure. Clinicians like to be able to treat diseases that resolve. When patients present with chronic conditions or an array of systemic symptoms that are hard to quantify, doctors do not have quick fixes to offer. These

© 2023 Scientific American



patients are often dismissed as malingerers or as suffering from a psychosomatic condition—and so it still is with long COVID.

Some patients have reported seeing doctors who want to help but lack the skills and bandwidth to do so. Early in the pandemic, staff at the Center for Post-COVID Care at Mount Sinai would spend hours with patients during their intake sessions. Compare that with the silo-based U.S. health-care system, which is designed to maximize efficiency: its basic building block is a 15-minute visit with a clinician. To treat long COVID effectively, then, Putrino thinks medicine needs more than just an infusion of interest and money. Additional funding, he says, will not "lead to a meaningful cultural change in the research and clinical world" until research centers start "actively involving people with these conditions" in the decision-making process.

The potential for transformation goes far beyond long COVID. Understanding what causes this condition might illuminate treatments for ME/CFS, tick-borne illness and other diseases that involve dysfunction of the immune system, many of which are on the rise. "I believe that understanding the pathogenesis of long COVID not only will help reveal parallel mechanisms for ME/CFS but also may hold a key to understanding autoimmune diseases, as many autoimmune diseases occur postinfection," Iwasaki says.

It is time for medical researchers to investigate these long-contested illnesses with the full force of science's power and for medical educators to train doctors in how to effectively *care* for chronically ill patients. If they do not, they will be failing not only this generation of patients but many millions more to come. SA

Meghan O'Rourke is editor of the *Yale Review.* Her essays, criticism and poetry have appeared in the *Atlantic,* the *New Yorker* and many other publications. Her latest book is *The Invisible Kingdom: Reimagining Chronic Illness* (Riverhead Books, 2022).

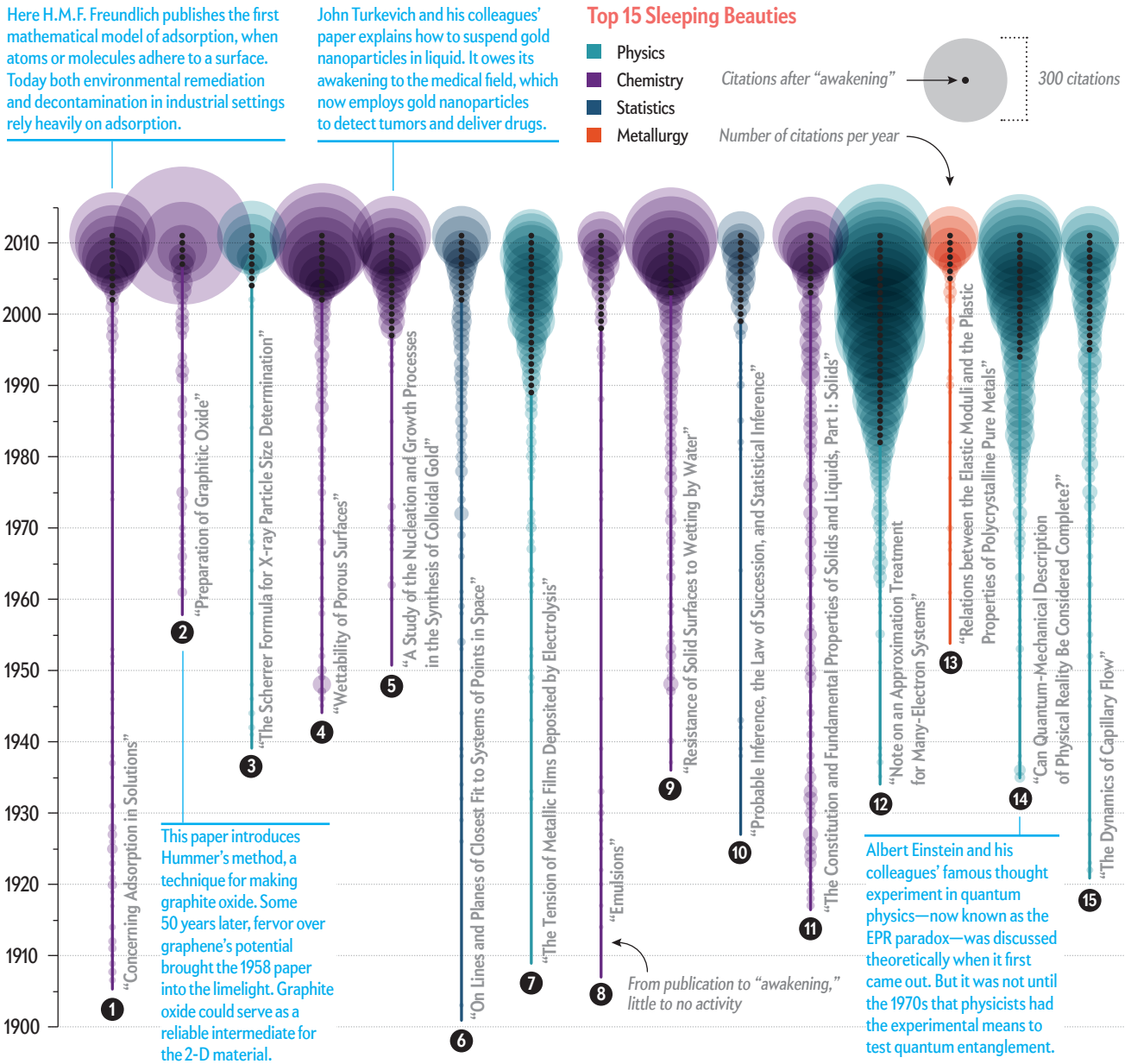
Illustration by Thomas Fuchs

© 2023 Scientific American

Sleeping Beauties of Science

Some of the best research can slumber for years

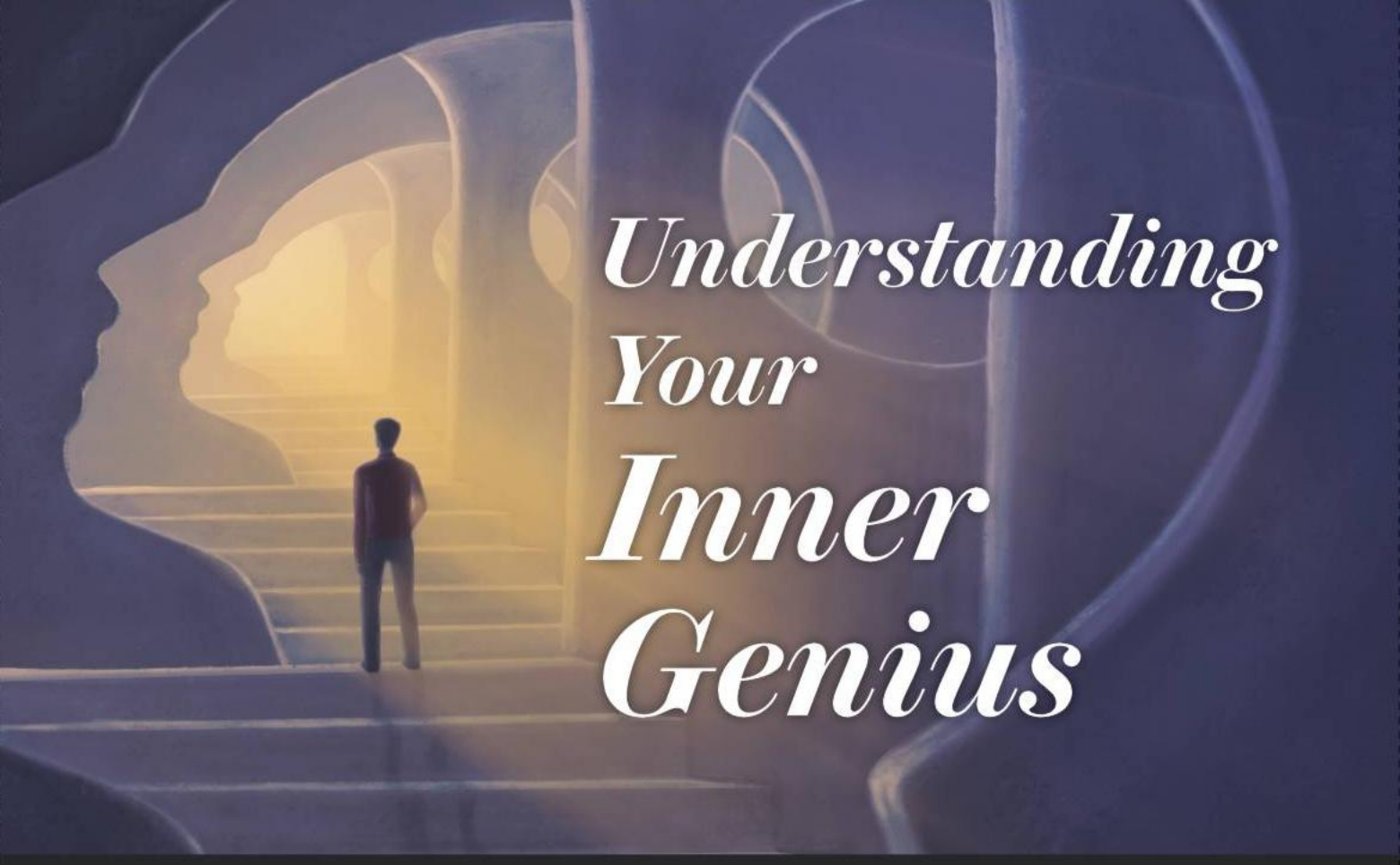
Discovery lies at the core of science, but rediscovery can be just as important. Indiana University Bloomington researchers combed through 22 million scientific papers published from about 1900 to 2010 and found, to their surprise, dozens of "Sleeping Beauties"—studies that sat dormant for years and then suddenly got noticed. The top finds—the ones that languished in obscurity the longest and later got the most intense attention from the scientific community—came from chemistry, physics and statistics. What tended to wake up these sleeping findings? Scientists from other disciplines, such as the medical field, in search of fresh insights, as well as the ability to test once theoretical postulations. Qing Ke, a data science professor at City University of Hong Kong, who worked on the project as a graduate student, says that most likely Sleeping Beauties will become even more common in the future because of the increasing availability and accessibility of scientific literature. *—Amber Williams*


Here H.M.F. Freundlich publishes the first mathematical model of adsorption, when atoms or molecules adhere to a surface. Today both environmental remediation and decontamination in industrial settings rely heavily on adsorption.
John Turkevich and his colleagues' paper explains how to suspend gold nanoparticles in liquid. It owes its awakening to the medical field, which now employs gold nanoparticles to detect tumors and deliver drugs.
Top 15 Sleeping Beauties
Physics
Chemistry
Statistics
Metallurgy
Citations after "awakening"
300 citations
Number of citations per year
2010
2000
1990
1980
1970
1960
1950
1940
1930
1920
1910
1900
1 "Concerning Adsorption in Solutions"
2 "Preparation of Graphitic Oxide"
3 "The Scherrer Formula for X-ray Particle Size Determination"
4 "Wettability of Porous Surfaces"
5 "A Study of the Nucleation and Growth Processes in the Synthesis of Colloidal Gold"
6 "On Lines and Planes of Closest Fit to Systems of Points in Space"
7 "The Tension of Metallic Films Deposited by Electrolysis"
8 "Emulsions"
9 "Resistance of Solid Surfaces to Wetting by Water"
10 "Probable Inference, the Law of Succession, and Statistical Inference"
11 "The Constitution and Fundamental Properties of Solids and Liquids, Part I: Solids"
12 "Note on an Approximation Treatment for Many-Electron Systems"
13 "Relations between the Elastic Moduli and the Plastic Properties of Polycrystalline Pure Metals"
14 "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?"
15 "The Dynamics of Capillary Flow"
This paper introduces Hummer's method, a technique for making graphite oxide. Some 50 years later, fervor over graphene's potential brought the 1958 paper into the limelight. Graphite oxide could serve as a reliable intermediate for the 2-D material.
From publication to "awakening," little to no activity
Albert Einstein and his colleagues' famous thought experiment in quantum physics—now known as the EPR paradox—was discussed theoretically when it first came out. But it was not until the 1970s that physicists had the experimental means to test quantum entanglement.

Source: "Defining and Identifying Sleeping Beauties in Science," by Qing Ke et al., in *Proceedings of the National Academy of Sciences USA*, Vol. 112, No. 24; June 16, 2015

Graphic by Tiffany Farrant-Gonzalez

© 2023 Scientific American



Understanding Your Inner Genius

NEW VIDEO COURSE

*Join instructor and Editor in Chief Laura Helmuth for an exploration of human intelligence in **Understanding Your Inner Genius**.*

This course includes 9 lectures that discuss the latest research in:

- **Enhancing brainpower**
- **Helping kids excel**
- **The origins of genius**
- **Learning throughout our lifespan**
- **Cultivating creative breakthroughs**

Go to:

www.TheGreatCourses.com/SciAm3

Brought to you by SCIENTIFIC AMERICAN + THE GREAT COURSES

Scientific American is a registered trademark of Springer Nature America, Inc.

Emboldening the Mind Since 1845

Unlimited Discoveries. Unlimited Knowledge.

Scan to learn more

Scientific American is a registered trademark of Springer Nature America, Inc.